Final report of ENGAGE Establishing Next Generation sequencing Ability for Genomic analysis in Europe

Hendriksen, Rene S.; Karlsmose Pedersen, Susanne; Leekitcharoenphon, Pimlapas; Malorny, Burkhard; Borowiak, Maria; Battisti, Antonio; Franco, Alessia; Alba, Patricia; Carfora, Virginia; Ricci, Antonia

Total number of authors:
33

Published in:
EFSA Supporting Publications

Link to article, DOI:
10.2903/sp.efsa.2018.EN-1431

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

EXTERNAL SCIENTIFIC REPORT

APPROVED: 8 June 2018

doi:10.2903/sp.efsa.2018.EN-1431

Final report of ENGAGE - Establishing Next Generation sequencing Ability for Genomic analysis in Europe

Rene S. Hendriksen[1], Susanne Karlsmose Pedersen[1], Pimlapas Leekitcharoenphon[1], Burkhard Malorny[2], Maria Borowiak[2], Antonio Battisti[3], Alessia Franco[3], Patricia Alba[3], Virginia Carfora[3], Antonia Ricci[4], Eleonora Mastrorilli[4], Carmen Losasso[4], Alessandra Longo[4], Sara Petrin[4], Lisa Barco[4], Tomasz Wołkowicz[5], Rafał Gierczyński[5], Katarzyna Zacharczuk[5], Natalia Wolaniuk[5], Dariusz Wasyl[6], Magdalena Zając[6], Kinga Wieczorek[6], Katarzyna Półtorak[6], Liljana Petrovska-Holmes[7], Rob Davies[7], Yue Tang[7], Kathie Grant[8], Anthony Underwood[8], Timothy Dallman[8], Anaïs Painset[8], Hassan Hartman[8], Ali Al-Shabib[8], and Lauren Cowley[8]

[1]Technical University of Denmark (DTU), National Food Institute, WHO Collaborating Centre for Antimicrobial Resistance in Foodborne Pathogens and Genomics, European Union Reference Laboratory for Antimicrobial Resistance (EURL-AR), Research Group for Genomic Epidemiology, Denmark
[2]German Federal Institute for Risk Assessment, BfR, Department for Biological Safety, Berlin, Germany
[3]Istituto Zooprofilattico Sperimentale del Lazio e della Toscana "M. Aleandri", Direzione Operativa Diagnostica Generale, National Reference Laboratory for Antimicrobial Resistance (NRL-AR), Rome, Italy
[4]Istituto Zooprofilattico Sperimentale delle Venezie, Food Safety Department, Legnaro (PD), Italy
[5]Department of Bacteriology and Biocontamination Control, National Institute of Public Health – National Institute of Hygiene, Warsaw, Poland
[6]National Veterinary Research Institute, Puławy, Poland
[7]Animal and Plant Health Agency (APHA), Bacteriology Department, Weybridge, Addlestone, Surrey, the United Kingdom (UK)
[8]Public Health England (PHE), Colindale, the United Kingdom (UK)

Abstract

The ENGAGE project (http://www.engage-europe.eu/) was a collaboration between eight institutions across Europe. The aim was to boost the scientific cooperation to use whole genome sequencing (WGS) analysis in food safety and public health protection. ENGAGE focused on *Escherichia coli* (commensal *E. coli*) and different *Salmonella* spp. serotypes. A total of 3,360 genomes, 778 and 2,582 of *E. coli* and *Salmonella,* respectively, were produced. These genomes were stored and shared among partners in a temporary repository to be submitted to the European Nucleotide Archive by the end of the project. Generated genomes were used for benchmarking exercises to assess the possibility of replacing conventional typing with WGS for outbreak investigation. For the analysed strains, the benchmarking exercises showed that SPAdes assembly performed better than Velvet and that, by using different bioinformatics tools, WGS *Salmonella* serotyping and antimicrobial resistance genes detection, were largely in concordance with phenotypic data. Discrepancies were related to sequence quality and phenotype misclassification rather than to limitations of the bioinformatics tools. All partners were able to infer the expected phylogeny for the *Salmonella* and *Campylobacter* isolates in benchmarking exercises. Two WGS proficiency tests (assessing different genomic quality markers) were conducted among partners with satisfactory results. Guidelines including available bioinformatics tools and standard operating procedures (wet and dry lab) were prepared and posted online. Workshops, training courses and twinning programmes were conducted. The training focused on

online, Galaxy-based, and command line bioinformatics tools. To reach out beyond ENGAGE, an e-learning course (17 videos) was developed and made available online. Several proof of concept projects were run and some outcomes published, e.g. the discovery of colistin resistance gene, *mcr*-5. Overall, the project showed that laboratories without previous WGS experience need a period of time to implement and perform WGS for foodborne pathogens routine analysis. All developed material will remain available on the ENGAGE website.

Key words: next generation sequencing, capacity building, training, sequencing data, benchmarking, e-learning, guidelines

Question number: EFSA-Q-2015-00733

Correspondence: biocontam@efsa.europa.eu

Suggested citation: Technical University of Denmark - National Food Institute; Istituto Zooprofilattico Sperimentale del Lazio e della Toscana; Federal Institute for Risk Assessment; National Institute of Public Health – National Institute of Hygiene; National Veterinary Research Institute; Public Health England; Animal and Plant Health Agency, and Istituto Zooprofilattico Sperimentale delle Venezie, 2018. Final report of ENGAGE - Establishing Next Generation sequencing Ability for Genomic analysis in Europe. EFSA supporting publication 2018:EN-1431. 252 pp. doi:10.2903/sp.efsa.2018.EN-1431

ISSN: 2397-8325

Summary

The project entitled "Establishing Next Generation sequencing Ability for Genomic analysis in Europe" - ENGAGE (http://www.engage-europe.eu/) established a collaboration to boost scientific cooperation between eight public health, food and veterinary institutions across the European Union (EU), in order to build and enhance the use of whole genome sequencing (WGS) and analysis in food safety and public health protection. The project partners were the Technical University of Denmark - National Food Institute (DTU Food) from Denmark, Istituto Zooprofilattico Sperimentale del Lazio e della Toscana (IZSLT) from Italy, Federal Institute for Risk Assessment (BfR) from Germany, National Institute of Public Health – National Institute of Hygiene (NIPH-NIH) from Poland, National Veterinary Research Institute (NVRI) from Poland, Istituto Zooprofilattico Sperimentale delle Venezie (IZSVe) from Italy, Public Health England (PHE) from the United Kingdom (UK), and the Animal and Plant Health Agency (APHA) from the UK.

The project implemented joint proof-of-concept WGS projects that focused on subtypes of *Escherichia coli* (*E. coli*) and *Salmonella* spp. All activities were framed around these projects with a number of specific tasks embedded in 11 work packages and steered via 5 phases.

Seven affiliated institutions (who did not receive funding but collaborated on certain activities) also joined the project: verotoxin producing *E. coli* (VTEC) EURL (Italy), *Salmonella* EURL (the Netherlands), *Listeria monocytogenes* EURL (France), Institute of Food Safety, Animal Health and Environment "BIOR" (Latvia), Laboratorio Central Veterinario-Sanidad Animal (Spain), the Norwegian Veterinary Institute (Norway) and the United States of America Food and Drug Administration (US FDA). A Code of Conduct (CoC) was signed between the consortium and the affiliated partners where relevant to ensure the protection of shared genomic data.

The ENGAGE consortium defined the criteria for the selection of strains and genomes to be included in the project: to include isolates from the nine most common *Salmonella* serotypes from both, human and food/animals infections, commensal *E. coli* as well as multidrug resistant (MDR)/extended spectrum beta-lactamase (ESBL) producing *Salmonella* and *E. coli* from the EU AMR monitoring programmes. It was also decided to keep the list flexible to target future emerging sub-types. Additional relevant and already available genomes were identified among the partners for both proof-of-concept projects and benchmarking activities.

DTU, IZSLT, BfR, NIPH-NIH, NVRI, and IZSVe participated in the initial twinning programmes. For these, partners brought their own strains to the hosting institute, DTU, for DNA extraction and purification, library preparation, WGS and analysis. In later twinning programmes, DTU and IZSVe visited PHE and APHA to learn how to use the PHE bioinformatics pipelines or to get acquainted with the IT infrastructure used.

Most of the institutes including DTU initiated the process of outsourcing WGS at the beginning of the project due to an approximately 50% cost reduction compared to running WGS in house. As a result of being part of ENGAGE, BfR and IZSVe obtained WGS platforms at the end of 2016, and IZSLT and NVRI at the beginning of 2017. Towards the end of the project period, all but one institute (NIPH-NIH) acquired sequencing platforms and established the ability to sequence isolates as well as to conduct bioinformatics analysis in-house, thus ceased outsourcing of sequencing.

A working space infrastructure was developed for the purpose of hosting produced genomes. The ENGAGE project facilitated the production of 3,360 genomes, of which 778 and 2,582 of *E. coli* and *Salmonella*, respectively. The 3,360 genomes were stored and shared among partners in a temporary repository (the working space) and were subsequently submitted to European Nucleotide Archive (ENA).

A list of online available bioinformatics tools and software was prepared in order to: (1) identify potential tools that could be used for benchmarking exercises focusing on *Salmonella* serotyping, identifying antimicrobial resistance genes, and assessing phylogeny of *Salmonella* and *Campylobacter,*

(2) identify potential tools to be used by the partners initiating WGS and lacking bioinformaticians, and (3) create an online repository of guidelines and standard operating procedures (SOPs) for building whole genome sequencing typing (WGST) capacity in laboratories outside the ENGAGE consortium.

A large amount of sequence data and a number of bioinformatics tools were tested within the six benchmarking exercises conducted during the project including: 1) *De novo* assembly tools (SPAdes 3.9 vs. Velvet 1.2), 2) Genotypic *Salmonella* serotype prediction, 3) Genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6,[1] 4) Genotypic detection of antimicrobial resistance (AMR) genes, 5) *Salmonella* Enteritidis phylogeny, and 6) *Campylobacter coli* phylogeny.

Two proficiency tests (PTs) organised within the Global Microbial Identifier (GMI) were executed once per year to evaluate the performance of the partner laboratories. The aim of the PTs was to assess the WGS quality during the period of the project. Two isolates of the six bacterial genus/species targeted, *L. monocytogenes*, *Klebsiella pneumoniae,* and *Campylobacter jejuni* (trial 2016), as well as *Salmonella enterica, E. coli,* and *Staphylococcus aureus* (trial 2017), were selected for the PTs and reference material was produced. Seven out of the eight consortium partners participated in testing one to three of the bacterial genus/species per PT trial. NIPH-NIH was not able to participate as they did not yet have an in house WGS platform. Summary reports of the conducted PTs are available on the ENGAGE website (www.engage-europe.eu).

The protocols and SOPs for DNA extraction, library preparation, and sequencing procedures and the list of available bioinformatics tools are publicly available on the ENGAGE website to maximise reach-out in boosting the scientific collaboration outside the project.

Two workshops and training courses were conducted. The first training course focused on basic bioinformatics analysis using the Center for Genomic Epidemiology (CGE) tools and basic Galaxy tools (platform used by PHE). This was held back-to-back with the annual workshop from the 10th to the 14th of October 2016 at the consortium institute, NIPH-NIH in Warsaw. The second training course and workshop were organized by IZSLT and took place from 23 to 27 October 2017 in Rome. This course dealt with the utility of the most relevant and commonly used bioinformatic software and tools from CGE or other institutions, executed in an UNIX environment and by command line.

Several joint proof-of-concept WGS projects that targeted various topics within ENGAGE were conducted, including:

- *mcr-1*-harbouring plasmids in *E. coli* in Denmark and their phylogenetic relationship with *mcr*-1 plasmids from other geographical regions (DTU Food)
- Phylogeny of *Salmonella* Paratyphi B variant Java (ST-28) harbouring *mcr-1* (BfR)
- Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in d-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B (BfR)
- VIM-1 producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany (BfR)
- *Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and antimicrobial resistance (AMR) genes (IZSLT)
- Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy (IZSLT)
- Genomic diversity of *Salmonella* Derby in different European countries (IZSVe)
- Whole genome characterization of *Salmonella* Napoli isolates spanning 2005-2015: a national issue of international interest (IZSVe)
- Comparative genomic analysis provides novel insights into the ecological success of the monophasic *Salmonella* Serovar 4,[5],12:i:- (IZSVe)
- WGS of rare and unrecognised *Salmonella* serovars (NVRI)
- *mcr-1* positive *E. coli* in Poland (NVRI)

[1] ISO/DIS 16140-6:2017 Microbiology of the food chain — Method validation —Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures.

As described in Section 3.10, various scientific contributions were published, presented at meetings, accepted or submitted to peer-reviewed scientific journals during the project period.

To maximize the outreach for capacity building, ENGAGE also produced an E-learning component consisting of 17 videos of a total length of 2 hours and 58 minutes describing the whole topic from basic conventional molecular sub-typing and the characterization of foodborne pathogens to the use of bioinformatics tools and the ENA submission tool (batch-upload) developed by CGE.

Six newsletters were issued on the ENGAGE website during the project period. These included information on management issues, the momentum of the project, the completed deliverables, results produced and related conclusions. In addition to the newsletters, information on the ENGAGE activities and the results of the serotyping benchmarking exercise were published as news on the GMI website (http://www.globalmicrobialidentifier.org/news-and-events/nyheder/Nyhed?id={3AD64758-03A2-4ACD-92ED-1DEEE427BA63} and http://www.globalmicrobialidentifier.org/news-and-events/nyheder/2016/09/providing-resources-and-guidance-for-next-generation-sequencing?id=e341eaa2-bad2-4afb-8d52-d14bb0c1e95c

All developed material will remain available on the ENGAGE website (http://www.engage-europe.eu/).

Table of contents

Introduction

Background and Terms of Reference as Provided by the Requestor

This grant was awarded by EFSA to: Technical University of Denmark, National Food Institute (DTU Food)

Beneficiaries: Technical University of Denmark - National Food Institute; Istituto Zooprofilattico Sperimentale del Lazio e della Toscana; Federal Institute for Risk Assessment; National Institute of Public Health – National Institute of Hygiene; National Veterinary Research Institute; Public Health England; Animal and Plant Health Agency, and Istituto Zooprofilattico Sperimentale delle Venezie

Grant title: ENGAGE - Establishing Next Generation sequencing Ability for Genomic analysis in Europe

Grant number: GP/EFSA/AFSCO/2015/01

Main objective of the call

The main objective of this call is to facilitate a scientific cooperation framework, the development and implementation of joint projects and the exchange of expertise and best practices in the fields within the Authority's mission. In particular, the action financed by the European Food Safety Authority (EFSA) grant to be awarded following the present call for proposals shall contribute to the objective of boosting scientific cooperation between scientists and research organisations with a competence in the development and validation of new approaches in the area of microbiological and chemical hazard assessment. It is of paramount importance to coordinate efforts between the food, veterinary and human health sectors in order to obtain maximum benefits from the use of WGS and read across methodologies for microbial and chemical food safety, respectively.

Specific objectives of the call

Making use of molecular approaches to identify and characterise microbial foodborne pathogens, specifically using WGS analysis, to enhance the understanding, the traceability and spread of the disease in humans that these bacteria populations may cause.

The molecular approach to identify and characterise microbial foodborne pathogens, specifically using WGS analysis, provides a golden opportunity to (i) explore the bacterial genetic diversity within and between compartments in the food chain; (ii) to assess the epidemiological relationship of isolates from different compartments; and (iii) to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans (e.g. virulence and antimicrobial resistance). The methodology is very promising, and the technology is still evolving quickly. However it is still unclear when and how this technology will be ready to be applied to routine activities and proof-of-concept projects for application in a public health context are needed.

Interpretation of the Terms of Reference

The specific objectives of this call focus at using WGS to identify and characterise foodborne pathogens as described in Section 1.1. The ENGAGE consortium opted also to focus on exchanging expertise, developing and providing training on WGS, providing consensus quality parameters on NGS outputs applied to WGS-based characterisation of bacterial pathogens, providing benchmarking exercises on bioinformatics tools, producing SOPs and training materials as described below.

1) To build a One Health, all sectors network to boost scientific cooperation

a. To achieve the One Health approach, we included partners from the three sectors: public health, food, and the veterinary sectors. Thus, we contacted the EURL-AR network, the European Commission (EC), Food and Waterborne Disease (FWD) network of the European Centre for Disease Prevention and Control (ECDC), and partners within the "COllaborative Management Platform for detection and Analyses of (Re-)emerging and foodborne outbreaks in Europe" (COMPARE; http://www.compare-europe.eu/) consortium working with bacteria.

2) To develop and implement joint proof-of-concept WGS projects on foodborne pathogens (FBP) to investigate genetic diversity, epidemiological links, successful clones, virulence and AMR of isolates from different compartments

a. Launching projects that focused on *E. coli* as, globally, *E. coli* is one of the most frequent bacterial species both in human and animals. Some *E. coli* strains are important human bacterial pathogens that cause a variety of infections including poorly understood foodborne infections as well as large outbreaks. Most of these projects focused on commensal *E. coli*, as these bacteria are subjected to high antimicrobial pressure and are interesting from the AMR point of view. Additionally, very little is known about clones and serovars circulating as serotyping is rather rarely used. The other selected target organism was *Salmonella* spp. as this is still one of the top FBP in EU. For both organisms, infectious strains resistant to cephalosporins and carbapenems persisting or emerging in the food chain have been reported.

b. Collecting strains to be whole genome sequenced and storing sequences temporarily in a central working space for comparative analysis. This also included genomes of strains that had already been sequenced by consortium partners.

3) To exchange expertise and best practices in WGS and bioinformatics analysis

a. Organising workshops, training courses and twinning programmes for consortium partners (self-supported network participants) and E-learning for the entire network for bioinformatics analysis of WGS data in order to disseminate knowledge, share best practices, SOPs and WGS analysis results.

4) To develop and appraise new bioinformatics tools used in microbial risk assessments of foodborne pathogens

a. Storing the data collected in a temporary working space, thus allowing the partners to individually i) investigate clonal diversity and epidemiological links, ii) search for virulence and AMR markers in major clones and iii) assess successful clones.

b. Establishing PT/QC guidelines to ensure reliable WGS data for reliable cross sector/cross country comparability.

c. Performing benchmarking exercises to assess the performance of different bioinformatics tools. We aimed to provide the results of such exercises to a broader audience during joint workshops.

Objectives

The objective of ENGAGE (Establishing Next Generation sequencing Ability for Genomic analysis in Europe) was to establish collaboration between the public health, food and veterinary sectors across the EU and to build capacity for WGS and bioinformatics analysis to ensure better food safety and public health protection.

The ENGAGE project boosted the scientific cooperation among leading food-, veterinary- and public health institutes across Europe by implementing joint proof-of-concept WGS projects to investigate: 1) bacteria evolution and genetic diversity driven by genetic exchange, selection, and clonal expansion; 2) epidemiological relationships between the target microorganisms; 3) identification of successful clones; and 4) microbial determinants of virulence, AMR, and other putative markers in various emerging 'epidemic' strains of commensal *E. coli* and *Salmonella* spp. that have the potential to have a strong impact in public health across Europe.

The ENGAGE project pursued the following specific objectives I – V within the format of a two-year program. Together these five specific project objectives shaped the overall framework of ENGAGE.

I. Joint proof-of-concept WGS projects targeting emerging commensal *E. coli* and *Salmonella* spp. Isolates were selected and collected by consortium partners within the first quarter of the project period.

WGS data was generated for joint analyses by the consortium partners and used for a number of scientific publications. A total of 3,360 *E. coli* and *Salmonella* spp. were sequenced throughout the project period. The strains originated from the daily routine diagnostics and the national/EU monitoring programs. Affiliated partners were invited to participate in the workshops and training courses of ENGAGE.

II. Creation of a temporary database for hosting genome data

The genome data of emerging commensal *E. coli* and *Salmonella* spp. isolates were temporarily stored in a secure working space (genome database) developed in the first quarter of the project period. Subsequently, all data was released to the European Nucleotide Archive (ENA) and EFSA, by the end of the project period. The consortium partners had full access to the working space including raw reads and a minimum set of metadata consisting of isolation source, host, host status, pathogenic organism, strain, subtype, collection date and country.

III. Benchmarking of bioinformatics analytic tools to assess the applicability and integration of WGS into public health

Currently, there are no international standardized guidelines on how to analyse WGS data. Thus, it was important to validate the bioinformatics methods being used to ensure that whatever tool was applied, the same epidemiological links, successful clones, virulence factors and AMR markers would be detected in the same data set. To address this, various commercial and online open access bioinformatics analytic tools were used in a comparative benchmarking analysis to ensure comparable results despite differences in the tools being used.

IV. To establish and conduct WGS proficiency tests

To ensure reliable WGS data from organizations generating WGS data, it was of paramount importance to establish and conduct WGS PTs regularly (e.g. at least once a year). In addition to the PTs, best practices guidelines, protocols and SOPs for quality assurance (QA) were developed, as well as reference materials which were made available. In the framework of GMI, the DTU and PHE were

already involved in setting up a PT scheme consisting of two components: 1a. DNA extraction, purification, library-preparation, and WGS from live cultures; 1b. WGS of pre-prepared DNA. ENGAGE decided to use the PT designed in the framework of the GMI PTs, using the reference material already available to avoid duplication of work and draw from the synergy between the two programmes. Overall, the PTs assessed the quality of the obtained genome sequences. The results of the PT trials were disseminated annually to ENGAGE partners through specific ENGAGE reports.

V. To boost scientific collaboration between scientists, clinicians, and risk assessors to exchange expertise and knowledge and build capacity in WGS and bioinformatics analysis

In the first year of the project, ENGAGE developed and organised training courses, twinning programmes and workshops for consortium and/or affiliated partners in order to promote knowledge transfer and expertise exchange and boost scientific collaboration in the area of WGS. In addition, E-learning modules were developed to build WGS capacity. The consortium partners discussed regularly, with participation of EFSA, the latest results through conference calls. The European Centre for Disease Prevention and Control (ECDC) participated as an observer in important meetings and discussions. In cases of emergency, *ad hoc* meetings would be autonomously arranged to disseminate urgent information.

Data and Methodologies

ENGAGE was planned as a two-year program divided into five phases that was carried out by 11 work packages (WP):

- WP1 Project Cooperation Framework
- WP2 Data Collection
- WP3 Whole Genome Sequencing
- WP4 Genome Database
- WP5 Analysis/Benchmarking
- WP6 Proficiency Testing
- WP7 Output/Outcome
- WP8 Training/E-Learning
- WP9 Workshops
- WP10 Dissemination
- WP11 Project Management

The first phase, the "Kick-off phase", involved three WPs facilitating the start-up process of the project (WP1 Project Cooperation Framework, WP2 Data Collection; and WP4 Genome Database).

The "Analytic phase" was one of the three operational phases and included WP3 Whole genome sequencing and WP5 Analysis/Benchmarking.

The "Proficiency testing phase" was the second operational phase and was related to WP6 Proficiency testing.

The "Boosting scientific collaboration phase" was the third operational phase and included WP7 Output/outcome, WP8 Training/E-learning and WP9 Workshops.

The "Administrative phase" included WP10 Dissemination and WP11 Project Management and continued throughout the project period.

Programme Phases and Work Packages

The *Kick-off phase* involved three work packages: WP1, WP2 and WP4. WP1, Project cooperation framework, provided a framework for the project with the aim of identifying and inviting affiliated partners including EURLs, EFSA, ECDC, relevant National Reference Laboratories (NRLs) and FWD network members. WP2 Data collection aimed to develop a sampling strategy and to identify additional relevant and already available genomes among the partners for both, proof-of-concept projects and benchmarking activities. WP4 Genome database was the third WP component of the Kick-off phase and the aim of this work package was to develop a working space at DTU to temporarily store collected genomes until submission to ENA.

The *Analytic phase* included WP3 and WP5 in which WP3 Whole Genome Sequencing included identification of already available strains for WGS, as well as identification and sequencing of strains from strain collections. We anticipated generating approximately 3,000 genomes of the target organisms, namely selected *E. coli* and *Salmonella* spp. strains, during the project period. This task would feed directly into performing sequence analyses and subsequently into ongoing and real-time analysis (WP5) of the generated genomes as benchmarking analysis reports and proof-of-concept projects. This phase continued throughout the project ending with all genomes submitted to ENA and EFSA.

The *Proficiency Testing phase* (WP6) explored the possibility for collaboration with already existing activities, e.g. GMI, to avoid duplications. Consequently, existing documents, protocols, etc. were re-used and modified to meet the requirements of a PT for the ENGAGE project. Based on the PT results, a quality assurance guideline[2] and a bioinformatic QC pipeline were developed. The QC pipeline was integrated into the CGE batch-upload tool[3]. In addition to the PT, WP6 also produced or provided relevant reference material of recent outbreak strains, strains harbouring emerging AMR genes, or other relevant strains of the target pathogens.

The *Boosting scientific collaboration phase* included WP7 in which a range of protocols and SOPs were developed to assist scientists in DNA extraction, library preparation and the execution of WGS and bioinformatics analysis. Based on identified suitable bioinformatics tools and software, a best practice guideline was developed with the purpose of indicating how to get started when embarking on WGS. Based on the developed guidance materials and previous experiences, the curricula for two three-day training courses running back-to-back with two 2-day workshops were developed, followed by invitations to workshops and training events (WP8 and WP9). The two workshops were organized to boost the scientific collaboration, and for the consortium partners to steer the project, ensuring that milestones were reached and deliverables developed. These meetings also provided a forum to disseminate the state-of-the-art knowledge and scientific presentations. Moreover, as part of WP8, twinning participants were identified and twinning programmes established and conducted during the project period, with most programmes running at the beginning of the project period. To ensure a proper outreach and sustainability of the developed material, E-learning lectures were developed based on training course presentations and posted online with open access. Developing the E-learning modules added to the sustainability of the project results.

The *Administrative phase* included administrative support to consortium partners, arranging conference calls, organizing kick-off-meeting, interim-meeting, and the final meeting with partners and EFSA, as well as reporting the progress and dissemination of news and results (WP10 and WP11). The administrative efforts also included maintaining the consortium information flow as a whole, coordinating activities with EFSA and INNUENDO[4], collaboration with INNUENDO in relation to their

[2] Available for download on the ENGAGE website; http://www.engage-europe.eu/resources/protocols-and-training

[3] https://cge.cbs.dtu.dk/services/cge/

[4] A novel cross-sectorial platform for the integration of genomics in surveillance of foodborne pathogens (http://www.innuendoweb.org/). The INNUENDO project has received funding from European Food Safety Authority (EFSA), grant agreement GP/EFSA/AFSCO/2015/01/CT2 (New approaches in identifying and characterizing microbial and chemical hazards) and from the Government of the Basque Country.

participation in a number of activities such as the benchmarking of serotyping tools, workshops and training courses.

Results

Establishment of the ENGAGE Cooperation

In January 2016, as part of WP1, Project cooperation framework, potential affiliated partners were identified and invited to participate in ENGAGE (not co-funded) by providing them with the opportunity to link with the project consortium and their activities, i.e. DTU Food reached out to relevant EURLs (EURLs of VTEC, *Salmonella* and *L. monocytogenes* (EURL-Lm)), the National Reference Laboratories (NRLs) of the EURL-AR network, ECDC, ECDC FWD network members, the US FDA, the US CDC, and WHO. WP1 also involved evaluating options to ensure compliance with the legal aspects of sharing data, strains and genomes including Non-disclosure agreements (NDAs), Memorandum of understanding (MoUs), Material transfer agreement (MTAs), or Codes of conduct (CoC).

The Legal department at DTU did not find it necessary to create a NDA, MoU or MTA to allow sharing of potential political sensitive genomic data and protect the Intellectual Property (IP). In February 2016, the legal department suggested to create a Code of Conduct (CoC) substituting the above documents.

In total, seven affiliated partners, which included the EURL-VTEC (Istituto Superiore di Sanità, ISS, Italy), EURL-*Salmonella* (the Dutch National Institute for Public Health and the Environment, RIVM, the Netherlands) and EURL-Lm (French Agency for Food, Environmental and Occupational Health Safety, ANSES, France), in addition to the Institute of Food Safety, Animal Health and Environment (BIOR, Latvia), the Laboratorio Central Veterinario-Sanidad Animal (Spain), the Norwegian Veterinary Institute (Norway), and the US Food and Drug Administration (USA), joined the project. The main reason why some of the contacted laboratories were reluctant to join ENGAGE was that they did not have genomic data to share nor did they yet have the capacity for WGS.

Working Space Infrastructure

Prior to the project launch, DTU as part of COMPARE had created a working space infrastructure which was further developed for the needs of ENGAGE for the purpose of hosting produced genomes. The working space allowed for upload/download of genomes either as singletons or in batches. Additionally, the working space allowed for submission of metadata via an excel sheet including key epidemiological markers such as date of isolation, origin (human, pig, etc.), source (stool, blood, etc.), geographical location of the sampling, as well as funding body, ownership, etc. This enabled the users to search for specific genomes for either benchmarking or other projects. The access to the working space was managed within WP4.

The working space allowed for uploads and downloads in batches as well as sharing data among partners. Some partners, i.e. PHE, BfR, IZSVe and APHA, who submitted their own genomes directly to ENA, included the ENGAGE project identifier to enable ENGAGE partners to locate the sequence data at ENA. In contrast, DTU downloaded genomes from NVRI, IZSLT and NIPH-NIH via the working space and submitted these genomes to ENA via the working space.

All the 3,360 isolates that have been sequenced during the project are available on the ENA website. The final dataset consists of 2,582 strains of *S. enterica* and 778 strains of *E. coli*. The Supplementary Table 1 (Annex A) references all the accession number for the sequences.

Data Collection

Early in the kick off phase and in a discussion with EFSA it was agreed that it would be preferable to expand the target to include commensal *E. coli* in general, and to have a strong focus on AMR for both target organisms, i.e. focusing on multidrug resistant (MDR)/extended spectrum beta-lactamase (ESBL) producing *Salmonella* and *E. coli* from the EU AMR monitoring programmes. In the selection criteria for strain and genome inclusion, efforts were made to avoid redundancies with the COMPARE (*S*. Typhimurium, monophasic *S*. Typhimurium, *S*. Enteritidis) and the INNUENDO (*S*. Enteritidis, VTEC) projects. However, as the scopes of COMPARE and INNUENDO differed from ENGAGE, it could be justified to include the above serovars in ENGAGE proof-of-concept and benchmarking projects.

A consensus consortium agreement was to focus on the nine most common *Salmonella* serotypes from both human and food/animals infections. This included the following serovars: *S.* Infantis, *S.* Kentucky, *S.* Stanley, *S.* Enteritidis, *S.* Paratyphi B var. java, *S.* Typhimurium, *S.* 4,[5],12:i:-/4,12,:i:- (monophasic *S*. Typhimurium) *S.* Newport, and *S.* Derby. Moreover, *S*. Napoli, a relatively uncommon *Salmonella* serovar in Europe, was taken into consideration as it is among the top serovars causing human infections in Italy, but the reservoir and transmission pathways are still unknown.

Additional focus was to target commensal *E. coli* as well as MDR/extended spectrum beta-lactamase (ESBL) *Salmonella* and *E. coli* from EU AMR monitoring programmes. The specific antimicrobial classes and drugs targeted within ENGAGE were the following: fluoroquinolones (ciprofloxacin), 3rd generation cephalosporins and cephamycins (cefotaxime, ceftazidime, cefoxitin), carbapenems (meropenem, imipenem), azithromycin, temocillin, tigecycline, and colistin. The selection process was kept flexible to allow targeting potential emerging sub-types but also fully susceptible strains. One of the laboratories also included some VTEC for specific studies (NIPH-NIH).

DTU, BfR, NVRI, and APHA provided strains of animal and food origin whereas PHE and NIPH-NIH provided human strains. IZSVe and IZSLT provided strains from all sectors. In summary, ENGAGE collected strains are representing the One Health approach. The provision of isolates was addressed in WP2 Data collection, which also contains a set of tasks. One of the consortium partners, DTU, is also the EURL for AMR, and as such supported the ENGAGE project with additional strains from other countries.

The information on the strains to be included in the project was captured from all partners in order to identify the strains available for sequencing (Appendix A). ENGAGE produced 3,360 genomes, 778 and 2,582 of *E. coli* and *Salmonella*, respectively. This included the following number of genomes per partner: DTU: 520, PHE: 500, APHA: 439, BfR: 382, NIPH-NIH: 320, NVRI: 368, IZSLT: 382, and IZSVe: 449. A summary of the 3,360 genomes, including ENA accession numbers, was provided in Supplementary Table 1 (Annex A).

Development of Protocols, Guidelines and SOPs for WGS and Analysis of WGS Data

Most activities in WP7 were related to facilitating the capacity building of laboratories which are in the process of implementing WGS. The WP includes guidelines and documents that support how to get started on performing WGS analysis and using bioinformatics tools. This included the following documents: the best practise guideline, "How to get started" (Appendix B), SOPs and protocols describing the DNA extraction, library preparation, and sequencing procedures using MiSeq, available bioinformatics tools (Appendix C and D) and a description about sequencing quality.[5] The WP also included writing all of the benchmarking reports allowing laboratories to identify the best tool for the analysis.

[5] Available on http://www.engage-europe.eu/resources/protocols-and-training

All produced protocols, SOPs, and guidelines are also publicly available on the ENGAGE website (http://www.engage-europe.eu/).

Whole Genome Sequencing

The consortium partners used a range of different sequencing technologies and platforms. This included Life Technology's Ion Torrent PGM and Illumina's Miseq, Nextseq and Hiseq platforms. DTU, BfR, IZSVe, NVRI and IZSLT have access to in-house Illumina Miseq. In addition, DTU had access to in-house Illumina Nextseq. PHE had access to in house Illumina Hiseq 2500 and APHA to two Illumina MiSeq/Roche 454/Mini-ION and Illumina NextSeq. Several partners - NIPH-NIH, NVRI, APHA, DTU, IZSLT - outsourced parts of the sequencing to subcontractors to save costs.

At the time when the project was initiated, PHE, APHA, and DTU had already established WGS capacity. The three institutes immediately initiated the in-house sequencing. In the early phase of the project, the remaining institutes, IZSLT, BfR, NIPH-NIH, NVRI, and IZSVe did not have an established WGS capacity or trained staff, and thus training of participants from these institutions through twinning was done in an early phase of the project. During the twinning at DTU, the participants brought their own strains for DNA extraction, library preparation and WGS including IZSVe: 9 strains of *S*. Derby and 25 monophasic variant of *Salmonella* Typhimurium; NVRI: 2 strains of *S*. Infantis, 6 strains of *Salmonella enterica* spp. *diarizonae*; BfR: 8 strains of *S*. Infantis; IZSLT: 7 strains of *S*. Infantis and 14 strains of *E.coli;* NIPH-NIH: 3 strains of *Salmonella enterica* subsp*. enterica* 1,4,[5],12:b:-, 2 strains of *Salmonella enterica* subsp*. enterica* 1,4,[5],12:i:-, 1 strain of *S.* Senftenberg and 1 strain of *S.* Paratyphi B var. Java. Subsequently, the genomes produced were analysed using the online CGE bioinformatic tools developed by DTU.

Later in 2016, most of the institutes, including DTU, initiated the process of outsourcing the WGS due to an approximate 50% cost reduction. This change was accepted by the financial unit at EFSA and the sequencing centres were included in the ENGAGE project as sub-contractors. In mid-2016, BfR and IZSVe developed in house capacity to perform WGS and bioinformatics analysis. Towards the end of 2016, IZSLT and NVRI also developed this capacity as a direct outcome of being part of ENGAGE. At the end of the project (January 2018) the only institute that was part of the ENGAGE consortium that had not established WGS capacity was NIPH-NIH and even this institute, as a direct consequence of their ENGAGE experience, is exploring (planning and cost calculation) the possibility of setting up its own Food and Waterborne Diseases WGS laboratory.

The appropriate and available bioinformatics pipelines and software for the WGS analysis and benchmarking were identified via Github, websites, and from the scientific literature. All partners used, and had access to, a number of available bioinformatics tools, either in house or online, for cluster analysis and detection of different epidemiological markers such as AMR genes.

Benchmarking Exercises

A list of bioinformatics tools and software for the analysis and benchmarking was generated and added to a list as the project progressed (Appendix D). The list presents an overview of bioinformatics tools available to analyse WGS data for different purposes and was intended to be used as a starting point to select and try out software available at the time of the study. As bioinformatics tools and software evolves with new techniques, it is advised to check for newly released software updates regularly. The list served three purposes: 1) listing the bioinformatics tools that could potentially be used for benchmarking exercises, 2) providing information for the partners initiating WGS and lacking bioinformatics capacities, and 3) summarizing in the ENGAGE website, the bioinformatics tools for the online repository of guidelines and SOPs for building WGS capacity (WP7) in laboratories outside the consortium.

Six benchmarking exercises were conducted and the reports summarised below present the details of the analyses. The benchmarking reports are available as appendices to this report and they have been posted on the ENGAGE website at http://www.engage-europe.eu/resources/benchmarking.

1. Benchmarking of *de novo* assembly tools: SPAdes 3.9 vs Velvet 1.2 (Appendix E): This benchmarking exercise was designed to compare two different assembly tools, SPAdes and Velvet. For the set of sequences analysed, the benchmarking showed that SPAdes generates longer contigs than Velvet. Furthermore, the accuracy of predicting the correct MLST in the sequenced *Salmonella* genomes was higher using SPAdes (100%) in comparison to using Velvet (94%).

2. Benchmarking of genotypic *Salmonella* serotype prediction (general) (Appendix F): General genotypic *Salmonella* serotype prediction of 786 different isolates (all strains from DTU were not part of ENGAGE project) covering 196 different serotypes using the MOST tool, SalmonellaTypeFinder, and the SeqSero stand-alone tool. During the project, this serotyping exercise was repeated with the inclusion of the SISTR tool developed by Public Health Agency Canada. The report presents the benchmarking of all four tools. Genotypic *Salmonella* serotype prediction of 786 different isolates covering 196 different serotypes using MOST, *Salmonella*TypeFinder 1.4, the SeqSero 1.2 stand-alone tool and SISTR v1.0.1, clearly demonstrated that serotyping of *Salmonella* spp. using NGS data is a very feasible option. The result of the exercise, for the set of sequences analysed, showed that the best bioinformatics tool, SISTR, achieved 88% (694 isolates out of 786) correlation with the conventional serotyping, followed by MOST and *Salmonella*TypeFinder with 85% (669 and 668 isolates respectively), whereas SeqSero had 65% correlation with the conventional serotyping, which is a conservative number considering that none of the isolates have been re-tested to ensure correct serotyping; moreover, different assembly tools were used by the participants which may also have an effect on the results. The 'no correlation' rate (miscorrelation, no prediction and ambiguous) was 12% for SISTR, 15% for MOST and *Salmonella*TypeFinder and 35% for SeqSero. The miscorrelation rates or cases where the tools predicted a different serotype than the expected were in the range of 3-4% for MOST, SeqSero and *Salmonella*TypeFinder and 8% for SISTR in this study. At least half of these miscorrelations are potentially due to mistakes in the conventional serotyping excluding the effect of using different assembly tools prior to benchmarking. Such a low miscorrelation rate would probably be hard to achieve for most laboratories performing conventional serotyping. It is recommended to perform classical serotyping of the isolates where the predictions from the tools disagree with the expected serotype. This is especially important for the isolates where all tools have identical miscorrelations. A higher miscorrelation rate was expected for the tested dataset as of the 786 tested isolates 500 were routine isolates provided by PHE, 208 diverse isolates provided by DTU (not sequenced as part of the ENGAGE project) and 78 rare isolates provided by APHA (i.e. antigenic formulas observed occasionally in routine laboratory diagnostics). The correlation of the serotype prediction with the conventional serotyping was > 90% with all tested tools (92-100%) for the 4 *Salmonella* serovars *S*. Typhimurium, *S*. Enteritidis, *S*. Kedougou and *S*. Kentucky. The results of this benchmarking exercise (Genotypic *Salmonella* serotype prediction) will be published in a peer-reviewed scientific journal.

3. Benchmarking of genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain — Method validation —Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures) (Appendix G): The main purpose of this benchmarking exercise was to perform an inter-laboratory study in order to evaluate a number of available bioinformatics tools for the *in silico* prediction of *Salmonella* serovars from raw WGS data. The setup of the study complied with the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain — Method validation —Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures). In this study, a total of 27 genomes (not part of ENGAGE project) were tested including 18 isolates of 6 *Salmonella* serovars: Enteritidis (n=3), Hadar (n=3), Infantis (n=3), monophasic Typhimurium (n=3), Typhimurium (n=3), Virchow (n=3); 5 non-target *Salmonella* serovars: Derby, Dublin,

Kentucky, Mbandaka, Stanley; 4 non-target genus, but of the same family (Enterobacteriaceae): *Citrobacter freundii*, *E. coli*, *Klebsiella pneumoniae*, *Shigella flexneri.* According to the current legislation (Reg. (EC) No 2160/2003 and subsequent implementing acts) on the control of *Salmonella* in poultry populations, a list of different serovars are considered relevant (called target serovars) and to control such serovars specific corrective measures must be implemented. All the other serovars are considered non-target and their control are based on different approaches. The results of this benchmarking study demonstrated that serotyping using WGS data is a promising option. The tools predicting the *Salmonella* serovars most optimally, in the current study, were SISTR, SeqSero, SalmonellaTypeFinder followed by MOST, resulting in a 96.3% correlation (SISTR (v1.0.1 and v0.3.6), SeqSero 1.0 (command line version), SalmonellaTypeFinder 1.3 (https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder/) and MOST) with the conventional serotyping for the set of sequences analysed. When analysing the data in accordance with ISO/DIS 16140-6:2017, the evaluation of results at species level showed to be within the acceptability limits for inclusivity and exclusivity as indicated in ISO/DIS 16140-6, but at serovar level they exceeded these limits. This latter was mainly caused by the fact that for 7 target strains, 3 tools could not identify the *Salmonella* serovar. Testing non-target strains additionally to target strains in such a study showed to be important as in three datasets *Citrobacter* was incorrectly identified as *Salmonella*. The choice of assembly tools and/or different options/parameter settings still needs some attention when using WGS for serotyping *Salmonella*. It is recommended to repeat conventional serotyping of the isolates where the predictions from the tools disagree with the expected serovar.

4. Benchmarking of genotypic detection of antimicrobial resistance (AMR) genes (Appendix H): The results obtained when detecting AMR genes in 164 *E. coli* and 125 *Salmonella* genomes (two of the 125 *Salmonella* genomes included in this study were not sequenced as part of ENGAGE project) using bioinformatics tools showed that predicting antimicrobial resistance using WGS is a feasible and realistic alternative to phenotypic susceptibility testing. Of the 4 tested tools (ResFinder 1.2, KmerResistance 2.1, SRST2 v0.1.7, PHE GeneFinder (no version available; tests performed on 01.02.2017)), the ResFinder 1.2 and the PHE GeneFinder tools provided the highest degrees of specificity, sensitivity, Matthew's Correlation Coefficient (MCC) and accuracy in the *Salmonella* dataset. ResFinder also provided the highest accuracy and MCC in predicting resistance in the *E. coli* genomes for the set of sequences analysed. However, in comparison to *Salmonella*, all 4 tested tools performed with a lower accuracy when testing *E. coli*; especially low accuracy was achieved in profiling β-lactam. This could be due to the inclusion of the *E. coli* containing upregulated chromosomal *ampC* mutation (mediating β-lactam resistance) in the dataset which none of the tools tested at the time could predict. As a result, the detection of *ampC* and chromosomal point mutations were included in the new, updated version of ResFinder 1.3. The results of this benchmarking exercise will be published in a peer-reviewed scientific journal.
5. Benchmarking for *Salmonella* Enteriditis phylogeny (Appendix I): The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools (online or command line), both to detect variants and to build a phylogeny based on the variants alignment detected for *S.* Enteritidis isolates. With the use of WGS, phylogeny is used as a method to investigate outbreaks and to perform surveillance among isolates that are genetically related. The results were compared using two main approaches: (1) Alignment and distance matrix comparison and (2) Topology of the tree: global topology, Robinson-Fould symmetric difference and percentage of edge similarity (number of branches in one tree that are present in another). The main conclusions of the study, for the set of sequences analysed, were that the SNP alignments generated by different methods showed similar results except in 3 cases when Snippy v3.0, CSIPhylogeny v1.4 method without heterozygote removal and with alignments produced by one customized pipeline (Centre 10) were used. The SNP alignments in these 3 cases showed a discrepancy between distantly related isolates The scores based on the topology demonstrate that most of the methods tested are able to retrieve the topology derived from the gold standard except for the method used by Centre 10 [VCF-kit v0.1.2 pheno tree nj]. During this benchmarking

we have identified that a key point in building a phylogeny based on the SNP-differences between isolates is the detection and filtering of the SNPs. Different tools/parameters can lead to the same topology but with a variable number of SNP- differences. This point is really important as new WGS investigation for outbreak based their case definition on a clustering approach (SNP-differences between isolates). Based on this benchmarking we recommend a minimum depth coverage for the SNPs detection > 10, a minimum mapping read quality of 30, and 90% consensus for the reads mapped at a position that differs from the reference. The best tools to build a tree from an alignment are maximum likelihood methods. Topology obtained using these methods produced trees with the best correlation between gold standard and the obtained phylogeny.

6. Benchmarking for *Campylobacter coli* phylogeny (Appendix J): The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools to detect variants and to build a phylogeny based on the variants alignment detected as a method to characterize outbreaks and perform surveillance of genetically related isolates. In this specific exercise, the participants were asked to take into account the possibility of recombination between isolates. The main conclusions from the study, for the set of sequences analysed, are that the methods used to generate the SNP alignment in the 11 different test sets showed similar results except for two centres where the comparisons of the distance matrix showed discrepancies between those isolates that are closely related. During this benchmarking we have identified that a key point in building a phylogeny where a recombination can occur is to link distance matrices and the phylogeny. The subtle topology of a closely related cluster is highly correlated to the presence of a recombination. We can also confirm with this benchmarking that, if the organism is likely to contain recombinations and discrepancies occur between phylogeny and epidemiology information, it is recommended to carry out a detection of recombination. This benchmarking shows that partners have used "gold standard" methods both for SNP detection and tree building. Filtering of SNPs has been properly carried out and most of the participants have used a maximum likelihood method to generate the phylogeny. It also demonstrated that, despite knowing the good practice to derive a phylogeny from WGS, phylogeny needs to be used with caution and can be only fully explained given support from other epidemiological data.

Overall, the benchmarking exercises showed that the tested bioinformatics tools were all performing as anticipated in serotyping *Salmonella*, identifying AMR genes and inferring phylogeny to *Salmonella* and *Campylobacter*. At this stage, it is not recommended to select only one single bioinformatics tool for the bioinformatics analysis since all tools are regularly updated and improved and all perform satisfactorily. All benchmarking exercises included the limitation that different assembly tools were used prior to testing the bioinformatics tools in question, potentially affecting the results.

Proof-of-Concept Projects

The ENGAGE project has been highly valuable for all ENGAGE consortium partners which also is reflected in the proof-of-concept projects executed as part of the ENGAGE project. The outcome of the ENGAGE project, including the selection of strains made by each ENGAGE consortium partner (covering the target organisms defined for WP3) and a summary of the proof-of-concept projects are presented below. Collaboration between several partners was facilitated and has led to valuable outcomes and in some cases to publications. Full descriptions of these projects are presented in Appendix K.

Feedback from DTU (Denmark)

The Technical University of Denmark, National Food Institute (DTU Food) was the coordinator of the ENGAGE consortium and had already, before the launch of the project, implemented whole genome sequencing (WGS) for routine analysis. At the point of the launch of the project, DTU Food already had access to the MiSeq (Illumina, San Diego, USA) and Ion Torrent (Life Technologies, Thermo

Fisher Scientific, Waltham, USA) platforms. In addition, DTU Food has developed a wide range of bioinformatics tools with open online access, i.e. the Center for Genomic Epidemiology tool-box (CGE). Even though DTU Food had already implemented WGS technology, the ENGAGE project contributed to and added value to several activities. Already prior to the initiation of the ENGAGE project, DTU Food was organizing and facilitating a proficiency test (PT) in the framework of GMI/WHO/COMPARE project. ENGAGE, however, ensured the momentum and keeping of deadlines of the PT during the project period. A working space, i.e. a sharing site for sequence data, was one of the tasks in the ENGAGE project in order to facilitate the sharing of genomes in a protected area. The working space was already scheduled and funded by COMPARE, but was not created until there was an immediate need for this in relation to ENGAGE. Similarly, E-learning had been discussed at DTU Food for some time but E-learning in relation to WGS had not been developed until the delivery of the videos for ENGAGE, the funding body of this activity. Basically, ENGAGE was the driver for these activities to be developed but none of the activities received double funding.

ENGAGE also provided the scene for a more targeted exercise to have bioinformatics tools benchmarked, which for DTU Food was valuable in the sense of providing knowledge about how CGE tools performed in comparison to competing tools.

Proof-of-concept project headed by DTU: *mcr-1*-harbouring plasmids in *Escherichia coli* in Denmark and their phylogenetic relationship with *mcr-1* plasmids from other geographical regions (Appendix K.1)

This proof-of-concept project was the result of collaboration between EURL-AR, DTU-Food and NRL-AR Italy. The aim of the project was to characterize *mcr-1* harbouring plasmids in *E. coli* in Denmark, also comparing them by phylogenetic analysis with those circulating in other countries.

For this purpose, the following isolates were selected by the NRL-AR for WGS: 115 *E. coli* isolates from food or animal origin from the Danish Integrated Antimicrobial Resistance Monitoring and Research Programme (DANMAP), out of which 50 were resistant to colistin (MIC ≥ 4 mg/L) and 65 were susceptible with an MIC of 2 mg/L. After screening for the presence of colistin resistance genes, 60 isolates were selected and whole genome sequenced (only 41 isolates were sequenced under the ENGAGE project).

Main results of this study are:

- 10 phenotypically susceptible isolates (MIC = 2 mg/L) harbor *mcr-1* (15% of the 65 colistin-susceptible isolates tested)
- 41 phenotypically resistant isolates (MIC > 2 mg/L) harbor *mcr-1* (82% of the 50 colistin-resistant isolates tested)
- *mcr-1* is integrated in IncX4 plasmids in 19 isolates (37% of the 41 *mcr-1*-positive isolates)
- One colistin-resistant isolate harbors no *mcr* genes but displays a point mutation in *pmrA/pmrB* two component system (*pmrB* V161G)
- Eight colistin-resistant isolates have no currently known mechanism of colistin resistance

Feedback from BfR (Germany)

Experts agree that next-generation sequencing is changing microbial genomics. Once implemented in food, veterinary and clinical laboratories, this technique allows continuous molecular surveillance of food-borne pathogens and improves consumer protection and food safety. However, implementing a new method is time-, cost- and labour-intensive.

The German Federal Institute for Risk Assessment (BfR) acquired the first Illumina sequencer at the end of 2014. Nevertheless, generation and evaluation of WGS data in a regular and routine manner

was considerably improved after training of staff in twinning courses within the ENGAGE project early 2016.

Through vibrant exchange of experiences between consortium partners, including the twinning at DTU (May 2016), as well as workshops in Warsaw (October 2016) and Rome (November 2017), the scientists at the BfR learned which protocols are most appropriate for library generation, sequencing and quality control and how simple bioinformatics analysis can be performed. Benchmarking exercises further improved the knowledge on availability and suitability of bioinformatics tools for sequencing data evaluation. The regular participation in proficiency tests ensured a consistently high quality of obtained data and results. In conclusion, participation in the ENGAGE project was the first step towards the future of real-time pathogen surveillance and outbreak investigation at the BfR.

In the two years of the ENGAGE project, the BfR sequenced more than 382 *Salmonella enterica* and *E. coli* isolates from animal, food and environmental sources using the in-house MiSeq and NextSeq (Illumina) sequencers. *Salmonella* isolates of the most common serovars as well as *E. coli* isolates were selected based on conspicuous phenotypic antimicrobial resistance profiles.

The BfR performed altogether three proof-of-concept projects to demonstrate the advantages of NGS use in microbial genomics.
In the first proof-of-concept project (see Appendix K.2) multidrug-resistant *d*-tartrate fermenting (*d*Ta+) *Salmonella enterica* subsp. *enterica* Paratyphi B (also known as variant Java) isolates from chicken with decreased colistin susceptibility were of particular interest. Initially, the focus was to provide new insights into the evolution and spread of the plasmid-mediated colistin resistance gene *mcr-1*. However, sequencing of *mcr-1* negative but colistin-resistant isolates finally led to the discovery of the novel transposon-associated and plasmid-encoded colistin-resistance gene *mcr-5,* which lead to another proof-of-concept project (see Appendix K.3).
In a third proof-of-concept project (see Appendix K.4), two recently identified carbapenem resistant isolates from the strain collection of the NRL for *Salmonella* at the BfR were analysed by WGS methods to get a deeper insight in the spread of carbapenemase encoding plasmids. In summary, findings obtained during the ENGAGE project were presented to the scientific community in several posters and presentations and resulted in four publications in scientific journals (references listed in project descriptions). An additional manuscript is in progress.

Proof-of-concept project headed by BfR: Phylogeny of *Salmonella* Paratyphi B variant Java (ST-28) harbouring *mcr-1* (Appendix K.2)

In 2015, plasmid-mediated colistin resistance was reported to be caused by a mobilized phosphoethanolamine transferase gene (*mcr-1*) in Enterobacteriaceae. Originally identified in *E. coli*, the gene was later reported in other Enterobacteriaceae including *Klebsiella* and *Salmonella.* Approximately 400 *S.* Paratyphi B *d*Ta+ isolates originating from food producing animals and food products received from 2006 to 2016 in the German NRL for *Salmonella* were selected for PCR screening for the presence of the *mcr-1* gene. Altogether 63 *mcr-1* positive *S.* Paratyphi B *d*Ta+ ST28 isolates were subjected to WGS. Results showed that the *mcr-1* gene was located on plasmids belonging to the IncHI2 or IncX4 replicon group. Strains isolated from 2008 to 2011 tend to carry the *mcr-1* gene on large multidrug-resistant IncHI2 plasmids and cluster in the phylogenetic tree, whereas strains isolated after 2011 mainly carry *mcr-1* on IncX4 plasmids and show a higher phylogenetic diversity.

Evaluation of data is still in progress. The complete genome sequence of the earliest *S.* Paratyphi B *d*Ta+ isolate harbouring *mcr-1* from the collection of the German NRL for *Salmonella* was published as an article in the journal Genome Announcement:

Borowiak *et al.* 2017: Complete genome sequence of *Salmonella enterica* subsp. *enterica* serovar Paratyphi B sequence type ST28 harboring *mcr-1*. Genome Announc 5:e00991-17

Proof-of-concept project headed by BfR: Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica serovar* Paratyphi B (Appendix K.3)

Plasmid-mediated colistin resistance is known to be caused by phosphoethanolamine transferases. At the beginning of the study, two mobilized colistin resistance genes termed *mcr-1* and *mcr-2* were known. In the study, a novel phosphoethanolamine transferase-like protein in 14 avian colistin-resistant, but *mcr-1* and *mcr-2* negative *Salmonella* Paratyphi B variant Java isolates was identified. The respective gene (1644 bp), further termed as *mcr-5*, is part of a 7,337 bp transposon of the Tn*3*-family usually located on multi-copy ColE-type plasmids. In one isolate an additional variant was detected which carries the *mcr-5* transposon in the bacterial chromosome. The incidence of *mcr-5* carrying colistin-resistant *S.* Paratyphi B *d*Ta+ isolates in the collection of the NRL *Salmonella* in Germany is low and restricted to 14 strains isolated between 2011 and 2013. Nevertheless, these findings suggest that mobilized colistin resistance genes might be more common than expected and raise concern on their real variety, their prevalence and distribution in Europe and other continents as well as their relevance in public health.

Results of the study were published:

Borowiak *et al.*: Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B. J Antimicrob Chemother. 2017; 72: 3317–3324

Proof-of-concept project headed by BfR: VIM-1 producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany (Appendix K.4)

Carbapenems are considered last-line clinical antibiotics for treating severe human infections caused by multidrug-resistant Gram-negative bacteria. In 2011, VIM-1 carbapenemase-producing *Salmonella enterica* subspec. *enterica* serovar Infantis (*S.* Infantis) and *E. coli* were isolated for the first time from livestock farms in Germany. Within this study, the first detection of a bla_{VIM}-harbouring *S.* Infantis recovered from food (minced pork meat) in 2015 was described. Mapping of NGS data to already published resistance plasmid sequences revealed that the plasmid harboured by the food isolate is 100% similar to the previously published plasmid sequence of isolate R27 from swine in 2011. This finding hints towards a link between the isolates and to a transmission of this plasmid or these *Salmonella* isolates from the primary production into the food chain. The occurrence of carbapenemase-producing Enterobacteria (CPE) in food and food-producing animals might bear a risk of getting colonized with CPEs and raises major public health concerns.

Results of the project were published:

Borowiak *et al.*: VIM-1-producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany. J Antimicrob Chemother 2017; 72(7):2131-2133

Borowiak *et al.* 2018. Complete genome sequence of a VIM-1-producing *Salmonella enterica* subsp. *enterica* serovar Infantis isolate derived from minced pork meat. Genome Announc 6:e00327-18

Feedback from IZSLT (Italy)

The participation of the General Diagnostic Department, Italian National Reference Laboratory for Antimicrobial Resistance (NRL-AR), Istituto Zooprofilattico Sperimentale del Lazio e della Toscana (IZSLT), as member of the ENGAGE consortium has given to the department the opportunity to improve their knowledge on applications of Next Generation Sequencing (NGS) technologies in the field of molecular epidemiology of foodborne diseases and related antimicrobial resistance, and to implement Whole Genome Sequencing (WGS) data analysis in the Department and in the NRL-AR

Italy, IZSLT laboratories. The ENGAGE project also represented for the NRL-AR IZSLT one of the main means to bring WGS to the attention of decision makers at the IZSLT and to proceed to the purchase of a MiSeq sequencer (Illumina).

The evaluation of the different NGS technologies on the market performed by NRL-AR IZSLT, based on principles and parameters of the Health Technology Assessment and used for the choice of the technology and the machine to be proposed for purchase, was shared within the ENGAGE consortium and the network of the NRLs at the 2017 EURL-AR Training Course. The training course was held at DTU Food Kgs. Lyngby on 26-28 September 2017 and the presentation was entitled "Advocacy - how to bring WGS to the attention of decision makers at the institute".

IZSLT participated in twinning programmes in 2016 hosted at DTU Food which provided the possibility to develop and improve the capacities on performing all the steps of WGS analysis: from the wet-lab component (DNA purification, library preparation, normalization and sequencing using a MiSeq machine, Illumina) to the bioinformatics analysis by using different web-based freeware tools (dry-lab component). Participation in these twinning opportunities and also in the two annual training courses held in Warsaw (October 2016) and in the one organised by the NRL-AR IZSLT itself in Rome (October 2017) has allowed to update and improve the capacities at IZSLT on performing the analysis of the raw reads directly retrieved from the sequencer, also using some tools (Trimmomatic, SPAdes, CGE Docker tools) from command line in a Linux environment and the analysis of the genomes assembled with the bioinformatic tools (e.g. Velvet, SPAdes). This training activity also helped to identify the pros and cons of using different tools and thus improving the attitudes for an informed decision on which tool to employ in different situations. In this regard, IZSLT has implemented routine protocols also for the identification and characterization of resistance genes and other genetic markers (e.g. virulence markers, fitness markers, housekeeping genes for population structure, etc.).

The acquisition of these new competences on WGS analysis represents a crucial turning point for the NRL-AR IZSLT in the area of surveillance of pathogens and related antimicrobial resistance (AMR) and on monitoring programmes and activities. In this regard, the laboratory is implementing an internal routine WGS protocol in the context of National and European AMR monitoring programmes for a basic evaluation of the quality of the raw reads, the assembly quality of the reads and the quality of the contigs. The main application of this protocol will be the detection of the many different molecular markers used for further characterization of pathogens and their accessory genome (e.g. housekeeping genes for Multilocus Sequence Typing, Plasmid Typing based on Replicons, PlasmidMLST, Integrons), virulence markers and molecular mechanisms leading to AMR (point mutations, acquired resistance genes), since WGS may potentially replace within a single laboratory procedure and analysis conducted by the use of *ad hoc* bioinformatic pipelines, a multitude of assays that so far had to be run separately by most of laboratories all over the world. This approach leads to a relevant improvement of the overall efficiency and helps in the optimisation of human and material resources.

Moreover, the participation of the NRL-AR IZSLT in the Proficiency tests organized within the two-years of the ENGAGE project has provided a robust and practical way to test the performance of the newly acquired capacities in WGS analysis and to identify those critical points in the process that would need further attention and monitoring. Comparing the results from IZSLT with those obtained by the other participants was also very useful, and seen as a challenge to learn and improve the application of new bioinformatics tools in order to reach a high degree of accuracy and a harmonised choice of quality parameters.

Proof-of-concept project headed by IZSLT: *Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and antimicrobial resistance (AMR) genes (Appendix K.5)

This project focused on *S.* Infantis, since this serotype is emerging worldwide, and it is one of the top five serovars involved in human infections in Europe. MDR-*S.* Infantis has increasingly been reported in Italy in the last years from food-animals and humans, and is also highly prevalent in several European countries in the broiler meat industry. The purpose of the project was to provide molecular

data based on WGS analysis in order to investigate similarities and differences within *Salmonella enterica* subsp. *enterica* serovar Infantis across Europe and across animal and food sources.

A total of N=229 *S.* Infantis isolated from humans (N=64), animal (N=115), and food/environment (N=50) sources in the frame of National Control Programmes and monitoring activities were analysed. Regarding the country of origin, 150 were isolated in Italy and 79 in other EU countries, including 20 from Ireland, 18 from Luxembourg, 26 from the Netherlands and 15 from Poland. In addition, a total of 153 sequenced genomes were added, including 34 provided by APHA (UK), 38 by BfR (Germany) and 80 by EURL-AR DTU-Food (Denmark).

To date, the IZSLT, in collaboration with the DTU, are doing further analyses on accessory genome (especially pESI-like plasmids) harboured by the isolates in the collection. Remarkably, specific markers of pESI-like megaplasmids, such as pESI backbone, *fim* and *K88* genes have been found in isolates from all the countries participating in the study. Further *in silico* analysis to confirm the presence of *S*. Infantis isolates harbouring the pESI-like megaplasmid are still ongoing.

SNPs-based phylogenetic and phylogeographic analysis by using the CSI phylogeny is still ongoing.

To date, a manuscript that includes part of the results of this proof-of-concept and with a working title of "*Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and AMR genes" will be ready to be submitted in a peer-reviewed scientific journal by the end of summer 2018.

Proof-of-concept project headed by IZSLT: Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy (Appendix K.6)

This project focused on the epidemiology of transferable colistin resistance mediated by *mcr* genes in food producing animals in Italy, since it is rapidly evolving across Europe and timely information on prevalence, trends and variants of *mcr*-positive isolates is needed to enhance surveillance and implement prevention and control measures. The purpose of the project was to investigate the molecular epidemiology of *mcr* genes and their genetic environment in commensal *E. coli,* Extended Spectrum Beta-Lactamase (ESBL)/AmpC-producing *E. coli*, and *Salmonella* spp. in different primary productions and foodstuffs of animal origin in Italy, collected over the last three-years (2014-2016) by the Italian NRL-AR in the frame of National Control Programmes and monitoring activities.

A total of 55 *E. coli* and 14 *Salmonella* of different serotypes phenotypically resistant to colistin and *mcr* positive by PCR, with the majority of them being also MDR including Extended-Spectrum Cephalosporin-resistant (ESC-R), were selected for WGS, along with 6 fully susceptible *E. coli* isolates included as negative control. All isolates had been collected from different primary productions (fattening turkey, broiler chicken, fattening pigs and calves). Additionally, three multidrug resistant (MDR) *S*. Infantis collected in 2016-2017, displaying a colistin MIC value ≥ 4 mg/L and *mcr*-1 positive by PCR with two of them being also ESC-R, were sequenced after the end of the ENGAGE project (December 2017) and then included in this project.

For a better interpretation of the results, the isolates were divided into two sub-groups according to the species, the serotype and the source, and deeply investigated:

1) 31 isolates collected from turkeys (28 *E. coli* and 3 *Salmonella enterica* isolates), 5 *E. coli* from pig samples and 6 *E. coli* from bovine samples.

2) 4 multidrug resistant (MDR) *S*. Infantis from broilers (N=3) and broiler meat samples (N=1), with two of them being also ESC-R.

To date, the main findings derived from this study were:

- The *E. coli* population harbouring colistin resistance mediated by *mcr* was highly heterogeneous.

- The *mcr* gene was detected in different *Salmonella* serovars (*S.* Typhimurium, *S.* Newport, *S.* Blockley and *S.* Infantis) circulating in Italy.

- The 4 MDR *S*. Infantis contained both pESI-like megaplasmids and IncX4 plasmids harbouring *mcr*-1.1.

Part of this study has been included in a scientific paper:

Alba P, Leekitcharoenphon P, Franco A, Feltrin F, Ianzano A, Caprioli A, Stravino F, Hendriksen R, Bortolaia V, Battisti A. Molecular epidemiology of *mcr*-encoded colistin resistance in Enterobacteriaceae from food-producing animals in Italy revealed through the EU harmonised antimicrobial resistance monitoring. Accepted for publication in the peer-reviewed scientific journal *Frontiers in Microbiology* (May 2018). doi: 10.3389/fmicb.2018.01217

To date, the IZSLT in collaboration with the DTU are preparing a manuscript that includes part of the results of this proof of concept, entitled "Colistin resistance mediated by *mcr*-1 in ESBL-producing, multidrug-resistant *Salmonella* Infantis in broiler chicken industry, Italy (2016-2017)" to be submitted in a peer-reviewed scientific journal.

Feedback from IZSVe (Italy)

The participation of the Italian Istituto Zooprofilattico Sperimentale delle Venezie (IZSVe) to the ENGAGE project has been a fruitful opportunity for building and enhancing the use of Whole Genome Sequencing and analysis of *Salmonella* to better address genotypic characterization in the everyday work of the Italian National Reference Laboratory (NRL) for Salmonellosis.

Before participating in ENGAGE, the Italian-NRL for Salmonella had no experience either in WGS data production or in WGS data analysis. At the end of the ENGAGE project, IZSVe has acquired autonomy in using the MiSeq instrument and has successfully sequenced, shared and uploaded to ENA 449 genomes of *Salmonella* strains.

Moreover, the participation in ENGAGE has allowed the IT-NRL for Salmonellosis to strengthen its networking capabilities and to facilitate transnational collaboration with international high-profile researchers and institutions across the EU. The expertise exchange with project partners and the opportunities of trainings and twinning conducted have been rewarding chances to harmonize procedures and pipelines and acquire best practices.

Indeed, the IT-NRL for Salmonellosis is now leading a number of scientific projects based on the use of WGS for *Salmonella* characterization and outbreaks analysis, thus increasing research excellence and enhancing national and international competitiveness.

Proof-of-concept project headed by IZSVe: Genomic diversity of *Salmonella* Derby in different European countries (Appendix K.7)

S. Derby is a serovar generally associated to pig chain; however, with a lower frequency, it is isolated also from other sources, such as turkeys, at least in specific geographical areas. Previous studies have demonstrated that different lineages are associated with different sources (Hayward *et al.*, doi: 10.1186/s12866-016-0628-4). This project was aimed at characterising *S.* Derby isolates from different sources and different lineages, as well as to identify specific genetic features that could explain host adaptation and persistence along the food chain in relation to specific sources. Also, human isolates were included in the investigation in order to infer the main sources of human infection associated to this serovar. Partners were asked to contribute to the project with 60-80 isolates from pigs, turkeys, humans and other species. Both isolates from animals and foodstuffs were considered, moreover, participants were asked to contribute also with human isolates. A total of 342 sequences of S. Derby were retrieved from 7 partners (IZSVe, BfR, DTU, NVRI, APHA, PHE and NIPH-NIH) whilst additional 87 strains were internally provided from PHE, for a grand total of 429 strains. A preliminary analysis of the sequences was carried out in the context of the twinning between IZSVe and PHE that took place in summer 2017. A subset of about 200 isolates was analysed according to the standard PHE internal pipeline, including: quality metrics checking, species identification, MLST based on the 7 housekeeping genes, MLST-guided serotyping prediction/serotyping from raw

sequencing reads, AMR *in-silico* characterization, eBURST Group (eBG) assignation and single nucleotide polymorphism (SNP) address assignation. The analysed isolates belonged to 9 different ST (39, 40, 71, 682, 683, 1326, 3135, 3857 and 3871), corresponding to three eBGs (57, 244 and 264). Isolates belonging to different eBGS were further investigated by multiple alignment in order to ascertain their level of similarity in relation to sources, geographical area of origin and other distinctive features. A selection of isolates was subject to *in-silico Salmonella* pathogenicity island (SPI) detection, confirming the massive presence of a panel of different SPIs. Further investigation will be conducted to ascertain the presence of SPI23 to test the hypothesis of its role in characterizing host adaptability.

This project has been a great opportunity of collaboration among ENGAGE partners in terms of sequences sharing as well as an opportunity for the proposing partner (IZSVe) to be trained by the PHE on the analytical approaches used on their routine. Data analysis for this project is, however, still ongoing.

Proof-of-concept project headed by IZSVe: Whole genome characterization of *Salmonella* Napoli isolates spanning 2005-2015: a national issue of international interest (Appendix K.8)

S. Napoli is among the top serovars causing human infections in Italy and the number of isolates belonging to this serovar has been rising since 2000. It is relatively uncommon in other European countries. Several outbreaks related to this serovar and associated to the consumption of Italian food products have been documented during the last years. *S.* Napoli is generally isolated from humans and environment, whereas strains from animals and food are quite rare. Several studies, using different approaches, tried to infer the sources of infection for this serovar, however, the reservoirs and transmission pathways are still partly unknown. Moreover, the interest toward this serovar is triggered also by epidemiological, clinical and molecular evidences revealing important similarities between *S.* Napoli and typhoidal serovars. In order to investigate this serovar and clarify its epidemiology, 157 *S.* Napoli isolates from three partners (IZSVe, BfR and DTU) were collected. At the time of the project start, no other partners had reported *S.* Napoli isolates. Isolates were sequenced and their metadata collected. Sequences obtained will be analysed to identify differences/overlaps among isolates from different sources as well as among isolates from humans and from other sources, to deepen insight into the potential virulence of this serovar as well as to identify the isolates closer to the human ones in order to infer probable sources of this emergent serovar.

This project has been a great opportunity to investigate *S.* Napoli, which is a peculiar serovar since, although limited in its geographical spread, it has a relevant epidemiological role. Data analysis for this project is however still ongoing.

Proof-of-concept project headed by IZSVe: A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic *Salmonella* Serovar 4,[5],12:i:- (Appendix K.9)

Over the past decades monophasic variant of *Salmonella* Typhimurium, *S.* 4,[5],12:i:- has been recognized as an emergent serovar for its rapid spread especially along the swine food chain. Although many studies have documented the ecological success of this serovar, few investigations have been conducted to explain this phenomenon from a genetic perspective.

In Italy, since 2011 *S.* Typhimurium 4,[5],12:i:- has ranked as the first serovar both from human and veterinary sources and there has been an increasing interest toward the identification of the putative markers, which could explain its epidemiological success and address the identification of effective control measures. Because of the relevance of this emergent serovar at national level, it was one of the main fields of research for IZSVe during the last years and when ENGAGE started IZSVe had collected a panel of *S.* Typhimurium 4,[5],12:i:- strains to investigate through WGS.

A comparative whole-genome analysis of 50 epidemiologically unrelated *S.* 4,[5],12:i:- isolates was performed. The isolates selected for the investigation were obtained from different sources over the last years. The main genetic trait shared by the investigated strains was represented by heavy metals tolerance gene cassettes. Functional studies were also performed to assess *S.* 4,[5],12:i:- capability to tolerate copper in the environment suggested, in order to ascertain that the acquisition of heavy metal

tolerance genes is useful for preventing the toxic effects of metals, thereby highlighting that this is a potential factor contributing to the success of this *Salmonella* serovar in farming environments. Moreover, phylogenetic analyses indicated a distinction among the investigated isolates based on the above described genetic traits, suggesting the involvement of different polymorphisms that give rise to multiple independent clones of *S.* 4,5,12:i:-.

Results of the project were published in Frontiers in Microbiology, section Food Microbiology: Mastrorilli E *et al.* (2018) A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic Salmonella Serovar 4,[5],12:i:-. Front. Microbiol. 9:715. doi: 10.3389/fmicb.2018.00715

This project was entirely conducted by IZSVe and ENGAGE provided an important technical and financial support to finalize this research.

Feedback from NIPH-NIH (Poland)

Despite the fact that WGS techniques are now widely available and were implemented in many European national laboratories, WGS has not been yet implemented at the Polish National Institute of Public Health – National Institute of Hygiene (NIPH NIH). Within the ENGAGE project the institute had the opportunity to apply WGS technology and this kind of advanced analysis on a large group of bacterial strains. During the project the first strains were sequenced in the whole genome scale. In total, 320 *Salmonella* strains from the public health sector in Poland were sequenced by outsourcing (7 strains sequenced at DTU during twinning, 167 strains sequenced in cooperation with PHE and 146 strains sequenced in cooperation with BioBank Lab, University of Lodz). All sequences were then submitted to the ENA database. Thanks to twinning, training courses, guidelines and E-learning lectures conducted during the project, knowledge was acquired on the sequencing process and, even more importantly, on sequence data analysis. Analysis of sequences enabled introduction to practical data analysis at NIPH-NIH particularly important for future epidemiological molecular investigations.

As a result, during the project, NIPH-NIH performed WGS analysis of e.g. monophasic *Salmonella* Typhimurium strains, *Salmonella* Enteritidis strains with decreased susceptibility to quinolones isolated in a hospital in Warsaw, as well as VTEC strains. Based on these experiences NIPH-NIH was able to conduct similar WSG analysis for *Salmonella* Enteritidis strains collected in a "point in time analysis" to investigate the epidemiological situation in Poland due to the outbreak connected with Polish eggs. During this study additional 93 *Salmonella* Enteritidis strains (79 from clinical samples and 14 from food samples) were sequenced and analysed using, among others, SNP address to assign the outbreak strains. As a result 9 strains (11.4%) isolated from the clinical samples were assigned as outbreak strains (6 from cluster 175 and 3 from cluster 360). From food samples, 4 strains were assigned as outbreak strains (all from cluster 175). These results indicated that the outbreak strains are prevalent in Poland. Further, more detailed analysis using WGS will be performed to investigate the outbreak deeply.

The intention was, based on the gained knowledge, to conduct similar future studies in the context of collaborative epidemiological outbreak investigations with other European laboratories. Additionally, based on the experiences obtained within both wet and dry lab, plans were made for implementation of WGS technology, initially as a project-based approach and ultimately for routine use.

Feedback from NVRI (Poland)

The ENGAGE project was crucial for implementing WGS technology at the Polish National Veterinary Research Institute (NVRI). NVRI has joined the consortium with neither experience nor WGS laboratory capacities. As partner in the project, this supported the successful application for governmental funding for setting up laboratory capacities, including designation of laboratory areas,

purchasing of Illumina MiSeq along with peripheral equipment and hiring the necessary personnel. By the end of the project NVRI was running a fully operational WGS laboratory.

The ENGAGE project, via trainings, workshops, and collaboration, provided opportunity for gaining skills, practical knowledge and experience in WGS and bioinformatic analyses of the results which have been verified by participation in intensive, multi-level proficiency testing. The benchmarking activities raised the awareness of multiple available tools and their performance in bioinformatics analyses.

The ENGAGE project has been highly useful to solve current diagnostic problems. Firstly, WGS was used in a real-life outbreak scenario. In autumn 2016, following the identification of a multi-country outbreak of *Salmonella* Enteritidis related to Polish eggs, thirteen strains originating from egg-producing farms identified in the epidemiological infection were sequenced (HiSeq, outsourcing). Following the acquisition of sequences, they were shared with EURL *Salmonella* (RIVM, NL). In return, several reference sequences (strains from outbreak confirmed human cases) were obtained for in-house comparison. The performed phylogenetic SNP tree analyses led to the following conclusion: 1) similarity to the reference sequences definitely confirmed Polish eggs as a source of the outbreak; 2) diversity of the tested strains indicate on possible multi-source infections of the flocks. Those conclusions were refereed by the General Veterinary Office (March 2017) that decided to increase the number of the strains to be sequenced. The sequencing performed later in 2017 (MiSeq, in-house) was conducted within reference activities on NRL *Salmonella* (outside ENGAGE project). The initial conclusions were confirmed and led to the concept of transport cages as a possible vector of multiple *Salmonella* strains introduction into numerous flocks located in 19 farms. All *S.* Enteritidis sequences were openly deposited in ENA.

Secondly, by identification of rare, exotic and highly similar *Salmonella* serovars, it helped to solve one of the burning diagnostic problems at NRL for *Salmonella*. *S.* Newport (6,8,20:e,h:1,2) and *S.* Bardo (8:e,h:1,2) as well as *S.* Senftenberg (1,3,19:g,[s],t:-) and *S.* Dessau (1,3,15,19:g,s,t:-) show similarities of antigenic formulas, thus producing difficulties in serotyping. WGS results provided evidence that isolates serotyped as *S.* Bardo and *S.* Dessau in fact belong to Newport and Senftenberg, respectively. Based on the genome comparison serovars Bardo and Dessau do not exist and they should be erased from Kauffman-White-Le Minor scheme. Furthermore, SeqSero failed to predict serovar based on gene identification, therefore current performance of bioinformatics tools should be taken with caution and improved accordingly. The findings will be presented to the scientific community during the International Symposium on *Salmonella* and Salmonellosis (submitted, September 2018) and are described in the proof-of-concept project headed by NVRI (Appendix K.10). WGS of other strains of incomplete antigen formula indicated that in fact they belong to well-known and epidemiologically relevant serovars (i.e. 9:-:- identified as *S.* Enteritidis). A number of strains representing *inter alia* other than *enterica* subspecies of *Salmonella enterica* were not perfectly identified with current bioinformatics tools.

Thirdly, the participation in ENGAGE research projects (i.e. on *S.* Derby, *mcr*-positive *E. coli*) gave the opportunity to improve quality of the research run at NVRI. *S.* Derby sequences (n = 56) were provided for the proof-of-concept project headed by IZSVe (Appendix K.7) whereas 80 *mcr-1* positive *E. coli* were used to analyze the epidemiological situation in Poland as an own proof-of-concept project (Appendix K.11).

Invaluable advantage of the project results from networking with the ENGAGE partners. The collaboration provided an opportunity for comparison of strains isolated across Europe. The conclusions drawn from WGS results are therefore more valuable and generalized. This cooperation will hopefully continue beyond the life of the project, for both research projects and for any future incidents and international outbreaks.

For the above reasons, NVRI already has and definitely will benefit from having been an active partner in the ENGAGE project.

Proof-of-concept project headed by NVRI: *WGS of rare and unrecognised* Salmonella *serovars* (Appendix K.10)

The aim of this project was to characterise *Salmonella* isolates causing difficulties in serotyping (incomplete or ambiguous antigenic structure and rare serovars from uncommon isolation sources) and to set up a database of reference sequences. For this purpose, the NVRI institute analysed 113 strains. The outcome of the project includes:

1/ confirmation that *Salmonella* Bardo should be actually included into *Salmonella* Newport; the isolates occurring occasionally in food chain in Poland are considerably diverse (SNP tree) and represent several sequence types (ST 31, ST 118, ST 166). Similar conclusions were drawn for *Salmonella* Dessau that should be included into *S.* Senftenberg (ST14, ST 210). Twenty-six strains complied with the selection criteria. Results of this study will be presented as a poster during International Symposium Salmonella and Salmonellosis (I3S) in September 2018.

2/ relevant gene identification confirmed serovar identification of several isolates of strange or uncommon origin i.e. European grass snake (*Salmonella* IIIb 28:z10:z, unknown ST; IIIb 38:r:z:[z57], ST 645; *Salmonella* Sunnycove, unknown ST), pet geko (*Salmonella* II 16:m,t:[z42], unknown ST), sheep (*Salmonella* IIIb 61:k:1,5, ST 432). Fifty strains complied with the criterion. Based on WGS they were classified into 33 *Salmonella* serovars.

3/ WGS of 37 strains with incomplete antigenic structure make it possible to classify them to the most common and *Salmonella* control programme relevant serovars i.e. *Salmonella* 1,4,5:-:- recognised as *Salmonella* Typhimurium, *Salmonella* 9:-:- (*S.* Enteritidis), *Salmonella* 6,7:-:1,5, (*S*. Choleraesuis, *S*. Thompson), *Salmonella 6,7:z10:-* (*S*. Mbandaka), *Salmonella* 35:-:- (*S*. Monschaui). Sequence types of the strains allowed their allocation in the common clones i.e. *S.* Enteritidis ST 11, and *S*. Mbandaka ST 413 occurring in food chain in Poland.

Proof-of-concept project headed by NVRI: *mcr-1* positive *E. coli* in Poland (Appendix K.11)

This project dealt with the characterisation of colistin-resistant (*mcr-1* positive) *E. coli* isolated from animals in Poland. The outcome of the project is the deep characterisation of *mcr*-1 positive *E. coli* along the food chain in Poland. Importantly, *mcr*-1 genes were found mostly (62%) in colistin susceptible strains (MIC = 2 mg/L) and mostly in turkey isolates. The gene was present in a quite diverse group of isolates (n = 80) belonging to numerous STs partly related to isolation source. Beta-lactam and quinolone resistance co-occurred, as well as chromosomally encoded polymyxins resistance mechanisms. Multiple plasmid replicons were identified, but the vast majority of strains carried IncX4 plasmid, suspected to be a *mcr*-1 carrier. Final results are presented in a manuscript on prevalence and characterization of *mcr*-1 positive *E. coli* isolated from food-producing animals in Poland that is planned to be submitted in June 2018.

Preliminary results of the project were reported (oral presentation) during the 2nd International Caparica Conference in Antibiotic Resistance, Caparica, Portugal, 12–15 June 2017.

Zając *et al.* 2017. Whole genome sequencing characterisation of *mcr-1* positive *Escherichia coli* isolated from turkeys and chickens, Proceedings of 2nd International Caparica Conference in Antibiotic Resistance, p 188.

Feedback from APHA (United Kingdom)

Molecular typing of pathogens is a priority, core Defra development that will incorporate sequencing and genomic analytical framework at the APHA (UK) for replacement of traditional typing methods with more powerful WGS based methods. Currently, this is being developed for three high priority statutory animal diseases – bovine TB, avian influenza and *Salmonella*. The results of the *Salmonella* molecular typing based on WGS will be used for operation of the *Salmonella* National Control

Programme and to compare *Salmonella* isolates with human isolates for human *Salmonella* Outbreak Control. Public Health England (PHE) has already moved to *Salmonella* WGS and the switch to WGS at APHA will allow harmonisation with PHE, which will increase efficiency and facilitate future outbreak detection and investigations nationally and also internationally. Within APHA, a *Salmonella* serotyping pipeline was established which benefits from several publicly available bioinformatics typing tools that were combined and tested in order to increase the reliability of the results. The pipeline outputs were tailored to fulfil the EU requirements for use of the White-Kauffmann-Le Minor (WKLM) serotyping scheme for veterinary *Salmonella* typing and also the APHA reference laboratory service requirements. The serotype is predicted by three different tools, MOST, SeqSero and SISTR that were part of the inter-laboratory benchmarking exercise carried out within ENGAGE (Appendix F) for which APHA sequenced 78 rare isolates (i.e. antigenic formulas observed occasionally in routine laboratory diagnostics). In addition, we have developed and tested a novel WGS-based method to differentiate *S.* Enteritidis, *S.* Typhimurium, *S.* Gallinarum and *S.* Pullorum live vaccine from field isolates. The results of the *Salmonella* serotyping benchmarking exercise and the performance of the APHA *Salmonella* pipeline will be published in peer reviewed journals.

The inter-laboratory benchmarking exercises and proficiency testing carried out within ENGAGE project enabled further evaluation of the performance of a number of bioinformatics tools. APHA sequenced further 124 *Salmonella* spp. genomes with different AMR determinants for inclusion in the benchmarking exercise '*Salmonella* Serotyping according to ISO' (Appendix H), contributed 10 genomes for the benchmarking exercise 'Salmonella Serotyping general' (Appendix G) and took part in the exercises '*S. Enteritidis* phylogeny' (Appendix I) and '*C. coli* phylogeny' (Appendix J). As part of the ENGAGE project, APHA was responsible for WP2, data collection, and for developing protocols/SOPs for DNA extraction, library preparation and sequencing. The ENGAGE project has been highly valuable to APHA in terms of providing opportunities to share WGS methods and practices with the scientists from other member states and in terms of developing long lasting collaborations.

Feedback from PHE (United Kingdom)

Public Health England (PHE) has been using WGS for routine surveillance and outbreak investigation of *Salmonella* since 2014 and has progressively introduced WGS methods for other gastrointestinal bacterial pathogens such as *E. coli*, *Shigella*, *Campylobacter* and *L. monocytogenes*. Since 2016, all of these pathogens have been routinely sequenced at PHE. The Gastrointestinal Reference Bacterial Unit (GBRU) works in conjunction with PHE's Genomic Services and Development Unit to sequence around 400 isolates per week with all isolates analysed with bioinformatics pipelines developed by the core bioinformatics group. The WGS analysis pipeline from PHE includes common components for all the gastro pathogens sequenced. This pipeline includes several components: quality and trimming analysis, species identification based on k-mer identification (KmerID https://github.com/phe-bioinformatics/kmerid), followed by MLST typing made by using MOST (Tewolde et al., 2016 - https://github.com/phe-bioinformatics/MOST) when public databases are available. Several specific components have been developed by the core bioinformatics group for serotyping (MOST and an altered version of SeqSero), stx subtyping for *E.coli* toxins (GeneFinder – in-house software, not publicly available), serotype prediction for *L. monocytgenes*. An AMR pipeline running with different databases for *Salmonella*, *E. coli* and *Campylobacter* spp. has been developed using a reads mapping approach. Finally, SNP typing is performed using a combination of software developed by the core bioinformatics unit and the Gastro Intestinal Reference Unit. SNP typing is performed by using the PHEnix SNP detection pipeline (http://phenix.readthedocs.io/en/latest/) and SnapperDB (Dallman et al., 2018 - https://github.com/phe-bioinformatics/snapperdb). Most of our in-house bioinformatics software is publicly available (https://github.com/phe-bioinformatics).

As part of the ENGAGE project, PHE was responsible, along with other consortium partners, for the bioinformatics benchmarking activities including AMR and phylogeny benchmarking. PHE contributed a total of 500 strains to the final ENGAGE collection and has provided further 39 isolates (only shared internally with partners) for the specific phylogeny benchmarking based on our outbreak detection (9

Campylobacter and 30 *Salmonella*). PHE also supplied isolates for specific projects like *S.* Derby. This experience was highly profitable as it enabled the PHE-developed AMR pipeline to be compared with other AMR pipelines using the same dataset and this has contributed to the validation of the PHE pipeline. A manuscript on the AMR benchmarking is in preparation in collaboration with DTU. Comparing PHE pipelines with other tools or gold standard methods through such benchmarking activities helps to validate that PHE software are performing well.

The ENGAGE project has also been highly useful in terms of collaborating with other partners. It provided the opportunity to work with other scientists from other member states and share WGS experiences and practice. The relationships developed will continue to give value beyond the life of the project. Together with other partners, PHE worked on the benchmarking activities to design, launch and write reports. PHE has also been extensively involved in the training session both with online tools and command line tools. This experience has been invaluable as the developed training material can be reused for training purpose within PHE. It has also provided a great opportunity for scientists to gain experience of designing and running such activities. Hosting two of the partners for twinning has also been a rewarding experience for both sides.

In terms of public health, the ENGAGE project has promoted the use of WGS across member states, enabling the comparison of strains between consortium partners which will enhance the investigation of any future incidents or outbreaks. Benchmarking tools by ENGAGE provides information on tools and helps ensure tools are comparable. Maximum benefit from WGS will be gained by the universal uptake of WGS and the sharing of results. Assisting other member states to do this has advantages for everyone involved, which is why PHE has and will benefit from being an active partner in the ENGAGE project.

Results of the Proficiency Testing

PT is a rather expensive activity due to the need for providing of the reference material such as closed genomes and DNA/cultures to the participants. Due to these circumstances, the original idea for including PT into the ENGAGE project was based on the synergy with the GMI PT testing and the opportunity to share the reference material and evaluation pipeline. This conflicted with the GMI selected target organisms for 2016 – *L. monocytogenes*, *K. pneumoniae* and *C. jejuni*. This, however, was not seen as an issue as the quality control is not dependent on the species but the ability to sequence. In 2017, the GMI aligned the target organisms ENGAGE to include *S. enterica*, *E. coli*, and *S. aureus*. Two isolates of the three genus/species per trial were selected and sent to the US company Microbiologics for lyophilisation. Subsequently, the produced reference material was sent to US FDA for closing the genomes to serve as reference genomes.

Seven out of the eight partners in ENGAGE participated in either the entire PT components or only targeted one of the species. NIPH-NIH was not able to participate as they did not yet have the in house WGS capacity. Overall, the PTs were useful exercises as they allowed ENGAGE consortium partners to assess the quality of their own data as well as to identify critical points for improvement. In general, all data analysed in the PT reports were satisfactory showing WGS proficiency among the seven partners (Appendices L and M).

Twinning Programmes, Training Courses and E-Learning

To ensure the early implementation of abilities to perform WGS and conduct bioinformatics analysis among the consortium partners, twinning opportunities were established in spring 2016. Twinning programmes aimed at providing exchange of expertise and best practices among the consortium members and also building capacity on WGS data production and analysis were conducted. The twinning programmes were provided by the experienced partners in the consortium: DTU, PHE and APHA. Twinning participants were identified on the basis of the declared needs of knowledge on WGS

data production or/and analysis of consortium partners. The first twinning programme was provided by DTU and consisted of a two weeks training in wet lab (DNA purification, library preparation, DNA sequencing on Illumina platform) and dry lab (WGS data analysis *via* CGE tools). A total of nine participants from six partner institutions were invited to DTU for one week (Table 1). The twinning was directed at less experienced in WGS participants. All participants brought their own bacterial strains for the sequencing (see Section 3.5). Details on participants are reported in Table 1.

Table 1: Participants of the first twinning programme provided by DTU

Consortium Partner	Participant	Type of twinning	Date
IZSVe	Sara Petrin	Wet lab/Dry lab	25/04/2016-04/05/2016
	Alessandra Longo	Wet lab/Dry lab	22/05/2016-01/06/2016
IZSLT	Patricia Alba	Wet lab/Dry lab	01/02/2016-13/02/2016
	Alessia Franco	Wet lab/Dry lab	13/06/2016-22/06/2016
	Fabiola Feltrin	Wet lab/Dry lab	13/06/2016-22/06/2016
BfR	Maria Borowiak	Wet lab/Dry lab	22/05/2016-01/06/2016
NVRI	Magdalena Zając	Wet lab/Dry lab	22/05/2016-01/06/2016
	Katarzyna Półtorak	Wet lab/Dry lab	13/06/2016-22/06/2016
NIPH-NIH	Tomasz Wołkowicz	Wet lab/Dry lab	13/06/2016-22/06/2016
EFSA	Beatriz Guerra Román	Dry lab	29/05/2017-31/05/2017

The second twinning programme was provided by PHE and APHA and consisted of a deep training in dry lab (data analysis and IT infrastructure management).
Two partners joined the programme, DTU and IZSVe, and the visiting researchers were Rolf Sommer Kaas and Eleonora Mastrorilli, respectively. One participant from DTU was invited to PHE and APHA for one week visit to discuss the UK implementation of IT infrastructure and workflows for conducting routine surveillance. One participant from IZSVe (Table 2) was invited to PHE for two weeks in the summer of 2017 to learn about UK bioinformatics tools used for surveillance and to boost collaboration on one of the proof-of-concept project (*S.* Derby). Details of this second twinning programme are reported in Table 2.

Table 2: Participants of the second twinning programme provided by PHE and APHA

Consortium Partner	Participant	Type of twinning	Date
DTU	Rolf Sommer Kaas	Dry lab	13/06/2016-17/06/2016
IZSVe	Eleonora Mastrorilli	Dry lab	17/07/2017-28/07/2017

The set objectives of the twinning programmes to address the needs of knowledge transfer in order to enable the institutions to enhance their networking and scientific capabilities were fully completed. Partner organizations with no prior experience have been provided with an opportunity to acquire expertise in WGS; procedures and bioinformatics pipelines have been harmonized.

Moreover, the consortium partners have strengthened their research excellence and increased scientific reputation and attractiveness. This enabled enhancement of collaborations amongst the twinning institutions that led to joint publications in peer reviewed journals and grant applications.

To also target outside users as well as the consortium partners, a training course focused on basic bioinformatics analysis using the CGE tools and to some degree Galaxy implemented tools (https://usegalaxy.org/), was conducted in October 2016. The training course was held back to back with the annual workshop from 12 to 14 October 2016 at the consortium institute NIPH-NIH, in Warsaw. A total of 18 participants from partner institutions and 8 participants from non-partner institutions participated in the course. In line with the collaboration with the INNUENDO project, the

coordinator of this project was invited to participate in the Workshop to present the project. Highlights from the two workshops held and the two training courses are available for download from the website.[6]

As a continuation of the first ENGAGE training course held in 2016 in Poland, a training course on "NGS analysis based on command line tools", organized by the Istituto Zooprofilattico Sperimentale del Lazio e della Toscana "M. Aleandri" (IZSLT, Italy), was held from 25 to 27 October 2017 at LAZIOCREA (Lazio Region), Rome, Italy. The aim of this course was to provide to the participants a basic knowledge of the tools used for the analysis independently of the web tools. At this regard, an overview of the LINUX systems and the use of tools based on command line for the NGS data analysis were proposed. This course was designed for users who had a basic knowledge in analysing NGS data using online tools and wished to acquire competence in more efficient NGS data analyses using command line tools. The event was attended by internal (ENGAGE Consortium) as well as external (non-ENGAGE Consortium) participants. A total of 19 participants with different background and level of experience in NGS data analysis took part in this course.[7] They represented a variety of institutions from different European countries (including the UK, Poland, Germany, Italy and The Netherlands).

To maximize the outreach for the capacity building, ENGAGE has produced an E-learning component consisting of 17 videos of a total length of 2 hours 58 minutes describing the topic from basic conventional molecular sub-typing to the use of CGE tools and batch upload. The video presentations available at http://podcast.llab.dtu.dk/index.php?id=292 were launched on 30 September 2017 and updated on 10 December 2017, and include:

1. General Principles in typing of Bacteria
2. Surveillance of Antimicrobial Resistance Using Whole Genome Sequencing
3. Application of Genomic Tools One Technology Take´s It All
4. Introduction to NextGeneration Sequencing (NGS)
5. *De Novo* Assembly, from Raw Reads to Contigs: Assembler Tool
6. Sequence quality of whole genome sequencing of bacteria
7. Species identification: KmerFinder tool description and applications
8. MLST Typing: MLST tool description and applications
9. cgMLSTFinder: core genome multilocus sequence typing tool
10. Resistance Gene Detection: ResFinder Tool description and applications
11. *Salmonella* Serotype Identification: SeqSero Tool Description and Application
12. *Salmonella* serotype identification: SalmonellaTypeFinder tool description and application
13. *E.coli* Serotype Identification: SerotypeFinder Tool Description and Application
14. Plasmid replicon identification and plasmid typing
15. Bacterial Analysis Pipeline Batch Upload
16. Phylogenetic Relatedness: CSIPhylogeny Tool Description and Application
17. Multipurpose detection of genetic markers MyDbFinder tool description and application

A Massive Open Online Course (MOOC) "Whole genome sequencing of bacterial genomes – tools and applications" at COURSERA was setup including the E-learning videos number 1-5, 7-8, 10-11 and 13-17. With just a few mouse clicks, students around the world can access this free online course and be introduced to whole genome sequencing techniques, which have revolutionized the way diseases are detected and outbreaks are investigated. The videos 6, 9 and 12 were filmed after the MOOC was launched and are therefore not included in the MOOC.

[6] http://www.engage-europe.eu/resources
[7] Complete list of participants is reported in the workshop highlights, see http://www.engage-europe.eu/resources

Students will learn the theory behind the methods and receive training in how to use various free online tools to analyse whole genome sequencing data in order to type bacteria and map the occurrence of resistance in the bacteria. There is continuous enrolment on the five-week course, which is taught in English. The course is aimed at people with an interest in the field – e.g. undergraduate or graduate students, laboratory technicians, researchers and people working in the food, health or veterinary sectors. Teaching takes place via an interactive textbook, which contains videos, quizzes and assignments. During the course students have the opportunity to meet in an online study group. Students should expect to spend one to two hours a week on their studies. People who complete the course are able to receive a course certificate for a fee of 50 USD. The course is offered through Coursera, which is an international provider of free E-learning courses. The full course description is available on Coursera's website at https://www.coursera.org/learn/wgs-bacteria. The course was launched on Coursera in September 2017 and by January 2018 more than 2000 learners were active.

Conducted Annual Workshops:

The first ENGAGE workshop was held at NIPH-NIH in Warsaw from 10 to 11 October 2016. The purpose of the workshop was to follow up on all activities with a focus on the benchmarking exercises and proof-of-concept projects as described above. A total of 22 participants attended the meeting, of which 15 were from partner institutions. The complete list of participants as well as the workshop minutes have been posted on the ENGAGE website.[8] A number of joint proof-of-concept WGS projects were planned during the meeting to target various target organisms selected within the WP 2. Further details of the proof-of-concept projects are described in Appendix K.

The second ENGAGE Workshop was held from 23 to 24 October 2017 at LAZIOCREA (Lazio Region), Rome, Italy, organized by the IZSLT, D. O. Diagnostica Generale, NRL-AR, Italy. The objectives of the workshop were to review the project progress, to communicate updates on project deliverables and to identify any potential emerging issues associated with the activities proposed in the Workshop held in Warsaw in 2016. A total of 24 participants attended the workshop, including the participants of the ENGAGE projects, representatives of EFSA and three non-ENGAGE participants, representatives of: EURL-VTEC (ISS, Italy), EURL-Salmonella (RIVM, The Netherlands), and INNUENDO consortium (EFSA co-funded project, coordinated by University of Helsinki, Finland; http://www.innuendoweb.org). The complete list of participants as well as the workshop minutes have been posted on the ENGAGE website.[8]

Dissemination of Project Progress to Consortium Partners and Project Results to the GMI Network

Project dissemination includes several components such as the general dissemination of information on the project activities during meetings, conferences, etc., but also via newsletters and GMI. The general dissemination activity of the project was an ongoing activity. As an example ENGAGE was mentioned several times during the WHO/PAHO Meeting on the Application of WHO Whole Genome Sequencing as a Tool to Strengthen FBD Surveillance and Response in Developing Countries, Washington DC, USA, 10-13 January 2017.

Six newsletters were issued on the ENGAGE website (http://www.engage-europe.eu) from May 2016 to October 2017, describing management issues, the momentum of the project and the information on the completed deliverables. In addition to the newsletters, information on the ENGAGE activities and the results of the serotyping benchmarking exercise were published on the GMI web site at http://www.globalmicrobialidentifier.org/news-and-events/nyheder/Nyhed?id={3AD64758-03A2-4ACD-92ED-1DEEE427BA63} and http://www.globalmicrobialidentifier.org/news-and-events/

[8] http://www.engage-europe.eu/resources

nyheder/2016/09/providing-resources-and-guidance-for-next-generation-sequencing?id=e341eaa2-bad2-4afb-8d52-d14bb0c1e95c).

The project management has encouraged the partners to timely submit produced genomes to the ENA repository. All acquired and generated genomes within ENGAGE were submitted and are already publicly available at ENA. All the sequences were also downloaded from ENA to the current WGS EFSA repository (Amazon S3 bucket).

Currently, five scientific papers containing results obtained during the ENGAGE project have been published in peer-reviewed journals. In addition, a number of draft publications are in progress and also posters and presentations containing ENGAGE results have been presented to the scientific community.

Journal articles

Alba P, Leekitcharoenphon P, Franco A, Feltrin F, Ianzano A, Caprioli A, Stravino F, Hendriksen R, Bortolaia V, Battisti A. Molecular epidemiology of *mcr*-encoded colistin resistance in Enterobacteriaceae from food-producing animals in Italy revealed through the EU harmonised antimicrobial resistance monitoring. Accepted for publication in the peer-reviewed scientific journal *Frontiers in Microbiology* (May 2018). doi: 10.3389/fmicb.2018.01217

Borowiak M, Hammerl JA, Fischer J, Szabo I, Malorny B. 2017. Complete genome sequence of *Salmonella enterica* subsp. *enterica* serovar Paratyphi B sequence type 28 harboring *mcr-1*. Genome Announc 5:e00991-17

Borowiak M, Szabo I, Baumann B, Junker E, Hammerl JA, Kaesbohrer A, Malorny B, Fischer J: VIM-1-producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany. J Antimicrob Chemother. 2017 Jul 1;72(7):2131-2133

Borowiak M, Fischer J, Hammerl JA, Hendriksen RS, Szabo I, Malorny B: Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B. J Antimicrob Chemother. 2017 72(12):3317-3324

Borowiak M, Fischer J, Baumann B, Hammerl JA, Szabo I, Malorny B. 2018. Complete genome sequence of a VIM-1-producing *Salmonella enterica* subsp. *enterica* serovar Infantis isolate derived from minced pork meat. Genome Announc 6:e00327-18

Mastrorilli E, Pietrucci D, Barco L, Ammendola S, Petrin S, Longo A, Mantovani C, Battistoni A, Ricci A, Desideri A and Losasso C (2018) A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic *Salmonella* Serovar 4,[5],12:i:-. Front. Microbiol. 9:715. doi: 10.3389/fmicb.2018.00715

Piekarska K, Wołkowicz T, Wolaniuk N, Zacharczuk K, Rzeczkowska M, Gierczyński R. Analysis of draft genome of three clinical strains of *Salmonella enterica* serotype Enteritidis ST11 with decreased susceptibility to ciprofloxacin. Med Dosw Mikrobiol. 2018; 70: 26-36.

Furthermore, the BfR is currently preparing a manuscript on the phylogeny of *mcr-1* harbouring *S.* Paratyphi B ST-28. The manuscript is planned to be submitted in 2018, and to date, the IZSLT, in collaboration with the DTU, are doing further analysis on accessory genome (especially pESI plasmids) harboured by *S.* Infantis isolates collected in the frame of the *S.* Infantis proof-of-concept project headed by IZSLT. Most likely, a manuscript that includes part of the results of this proof-of-concept and with a working title of "*Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and AMR genes" will be ready to be submitted in a peer-reviewed scientific journal by the end of summer 2018.

Moreover, the IZSLT in collaboration with the DTU are currently preparing a manuscript entitled "Colistin resistance mediated by *mcr*-1 in ESBL-producing, multidrug-resistant *Salmonella* Infantis in broiler chicken industry, Italy (2016-2017)" to be submitted in a peer-reviewed scientific journal, that includes part of the results of the proof-of-concept project "Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy" (Appendix K.6).

A publication based on the AMR benchmarking report (Appendix H) is in preparation with the tentative title: Benchmarking Antimicrobial Resistance tools using bacterial whole genome data.

Conference communications

Oral presentations:

Borowiak M. 2017: Phylogeny of *Salmonella* Paratyphi B variant Java harbouring *mcr-1*. ECCMID 2017, Vienna

Zając M, Sztromwasser P, Wasyl D, Hoszowski A. 2017. Whole genome sequencing characterisation of *mcr-1* positive *Escherichia coli* isolated from turkeys and chickens, Proceedings of 2nd International Caparica Conference in Antibiotic Resistance, p 188

Posters:

Fischer J, Borowiak M, Baumann B, Szabo I, Malorny B. 2017: Whole genome sequencing analysis of multidrug-resistant *Salmonella* Infantis isolates circulating in the German food-production chain. ECCMID, 2017, Vienna.

Borowiak M. 2017: Transposon-associated phosphoethanolamine transferase gene (*mcr-5*) is responsible for mobilized colistin resistance in *Salmonella* Paratyphi B variant Java (application for the poster competition for PhD students as part of the EFSA EU-FORA Fellowship programme).

Alba P, Feltrin F, Iurescia M, Amoruso R, Donati V, Caprioli A, Leekitcharoenphon P, Hendriksen RS, Battisti A, Franco A. Transferable colistin resistance mediated by the *mcr*- 1 gene is widespread among *Escherichia coli* and is emerging in *Salmonella* in the Italian fattening turkey industry. 26th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID 2016). 9-12 April, Amsterdam, the Netherlands.

Alba P., Feltrin F., Iurescia M., Amoruso R., Donati V., Caprioli A., Leekitcharoenphon P., Hendriksen R., Franco A., Battisti A. Transferable colistin resistance mediated by the *mcr*-1 gene is widespread among *Escherichia coli* and is emerging in *Salmonella* in the Italian fattening turkey industry. 2018. 18th International Symposium of the World Association of Veterinary Laboratory Diagnosticians. Sorrento (Italy).

Conclusions

The project entitled "Establishing Next Generation sequencing Ability for Genomic analysis in Europe" (ENGAGE, http://www.engage-europe.eu/) progressed as planned and has created much capacity and boosted the scientific cooperation to build and enhance the use of WGS and bioinformatics analysis in food safety and public health protection among the consortium partners, which are eight public health, food and veterinary institutions across the EU. Seven additional affiliated institutions (not co-funded by the project) which included the EURLs, EU NRLs and other major international organizations also joined and benefitted from the project. The project implemented benchmarking activities and joint proof-of-concept WGS projects that focused on subtypes of *E. coli* and *Salmonella* spp. All consortium activities were framed around these benchmarking activities and projects with a number of specific tasks embedded in the work packages.

The ENGAGE consortium selected isolates from partner institutions focusing on the nine most common *Salmonella* spp. serotypes including both human and food/animals infections, as well as on commensal *E. coli* and on MDR/ESBL-producing *Salmonella* spp. and *E. coli* from EU AMR monitoring programmes. Most of the institutes initiated the process of outsourcing the WGS due to an approximately 50% cost reduction compared to running WGS in house. Four consortium partners acquired WGS platforms during the project period, of which two as a direct result of being part of ENGAGE. In addition, all consortium partners established the ability to sequence isolates and conduct bioinformatics analysis.

A total of 3,360 genomes were produced, 778 and 2,582 of *E. coli* and *Salmonella* spp., respectively. All sequenced genomes were shared among consortium partners in the created ENGAGE working space and subsequently submitted to ENA. From there, they were downloaded to the current EFSA WGS repository.

Several of the sequenced genomes were included in six benchmarking exercises to assess the utility and performance of a number of bioinformatics tools, and for the proof-of-concept projects. The following six benchmarking exercises were conducted during the project: 1) *De novo* assembly tools: SPAdes 3.9 vs Velvet 1.2; 2) Genotypic *Salmonella* serotype prediction; 3) Genotypic *Salmonella* serotype prediction complying to the Draft International Standard 16140-6; 4) Genotypic detection of AMR genes; 5) *Salmonella* Enteritidis phylogeny, and 6) *Campylobacter coli* phylogeny. Overall, the benchmarking exercises showed that all tested bioinformatics tools performed in accordance with the expected phenotypic results in relation to serotyping of *Salmonella* and identification of AMR genes, as well as in relation to inferring phylogeny for *Salmonella* and *Campylobacter*. The benchmarking exercises were performed with the limitation that different assembly tools had been used prior to testing the bioinformatics tools in question. Differences in the quality of the sequences included in the data set analysed and assembly tool employed may have had an effect on the results of the bioinformatics tools tested. Additionally, although the overall performance of the tools used for genotypic *Salmonella* serotyping prediction may have been underestimated by some choices made in the study design, they were deemed as very good. Indeed, the inclusion of uncommon serovars, and possible mistakes/inconsistencies of the original conventional serotyping results in the set of *Salmonella* spp. used for the "genomic serotyping", may have been a source of discordant results between the *in silico* serotype prediction and the conventional serotyping method.

To maximise reach-out in terms of boosting the scientific collaboration outside of the project, a series of guidelines or protocols have been developed for DNA extraction, library preparation, and sequencing procedures as well as a list of available bioinformatics tools. In addition, series of videos were produced as an E-learning component with open access and were posted alongside the guidelines and protocols on the created website. The guidelines and on-line course represent a good overview of the initial steps for applying WGS technology in laboratories performing surveillance and research.

Six partners participated in a twinning programme that facilitated sequencing and analysis of 'own' strains. Furthermore, two workshops and two training courses were held enabling a more in-depth and comprehensive training on the use of bioinformatics tools. The presence of participants from different institutions with different backgrounds and experience in the field of NGS, optimised knowledge exchange between all partners. Subsequently, the quality of sequencing by seven partners was evaluated via participation in developed PT trials where all performed satisfactorily. The PT highlighted the importance of good practices at every step of the end to end WGS process. Moreover, as the results of the benchmarking exercises indicate, the choice of protocols and tools, even at genome assembly level, represents a crucial aspect of the reliability of the output of WGS analysis.

A number of joint proof-of-concept WGS projects targeting various topics within ENGAGE were conducted. These were based on data from strains sequenced during the ENGAGE project, focusing on the *Salmonella* serovars Infantis, Derby, Napoli, and monophasic variant of Typhimurium, rare and

monophasic *Salmonella* spp., AMR genes *mcr-1*, *mcr-5*, and bla_{VIM-1}, and phylogeny of *Salmonella* Paratyphi B var. Java. At the conclusion of the ENGAGE project (January 2018), the proof-of-concept scientific contributions were either published, or accepted or submitted to peer-reviewed scientific journals.

The deliverables and output generated within ENGAGE and hosted on the website provide sufficient guidance on implementation of WGS, should other European institutes, national authorities, official and reference laboratories choose to do this. Knowledge has been exchanged between the partner institutions and affiliated partners. Furthermore, consortium partners could act as facilitators and as disseminators of knowledge in their respective countries for laboratories intending to start working on WGS. Knowledge gained through this project has provided the ability to input into EFSA initiatives on assessment of the implementation of WGS in molecular surveillance of food borne pathogens. As part of the project period, ENGAGE has facilitated the introduction of fast-response, high throughput WGS and analysis methodology into EU institutions without previous experience in WGS. Such rapid implementation supports the feasibility of an EU-wide implementation of WGS that will facilitate future outbreak detection and investigation, identification of emerging strains with enhanced epidemic potential and epidemiological analyses. While maintaining the EU *Salmonella* control and hygiene regulations requirements for use of the conventional serotyping scheme for veterinary *Salmonella* typing, the close agreement between phenotypic and WGS-based serotyping of *Salmonella* spp. demonstrated that WGS-based serotyping to this standard is feasible and that the most relevant serovars were correctly identified, including the six major, regulated serovars.

Furthermore, the results of the benchmarking of AMR bioinformatics tools suggest that if chromosomal point mutations were to be included in the WGS-based AMR predictive bioinformatics tools for identification of AMR genes, phenotypic surveillance could in the future be supported by or replaced with the application of these tools.

Currently, more than 2000 specific AMR genes and multiple gene variants of AMR genes encoding the same resistance phenotype or efflux pump mechanisms that may be associated with increased risk, and the genetic elements associated with transferable resistance, can be identified by the tested bioinformatics tools.

Benchmarking of bioinformatics tools to detect sequence variants and to build a phylogeny based on variants alignment for *Salmonella* and *Campylobacter* further supported the use of WGS as a method to characterize outbreak strains and perform surveillance among isolates that are genetically related. However, as in standard methods, there is a potential inter-laboratory variability intrinsic in WGS methods, and as such differences may be associated with operational or data errors rather than reflecting true biological differences and characteristics it is essential for robust investigations that all apparently discrepant data are rigorously checked. At this stage, ENGAGE cannot recommend one single bioinformatics tool since all tools appear to perform satisfactorily and are being regularly updated and improved. Observed lower performance of some tools was often due to phenotypic misclassifications or limitations in quality of the sequences and different assembly tools being used prior to testing the bioinformatics tools in question and potentially affecting the results.

In the project period, ENGAGE has shown that it is possible to implement WGS and the use of bioinformatics tools in laboratories without any prior knowledge of WGS, and that other countries can be supported to do this through partnerships. In addition, ENGAGE has showed that some current phenotypic methodologies, e.g. *Salmonella* serotyping, could in the future be replaced by WGS and the use of bioinformatics tools. The ENGAGE project was successful on many levels both in terms of boosting WGS and analysis capacity and capability across Europe but also in demonstrating advantages of having genome data sets from different sources and different countries for validation and benchmarking exercises as well as investigative analyses. To date there has been little benchmarking of bioinformatics tools for microbial genome analysis and this project has contributed significantly to this which is beneficial to all who use such tools. A limitation to move the WGS

technology forward in zoonoses surveillance and food safety is likely due to lack of funding at institutions. Consequently, there is a risk of not being able to meet the future requirements in diagnostics and surveillance. We recommend, either at national or EU level, to provide more funding initiatives for the implementation of WGS for food safety in laboratories without current capacity.

Additional Supporting Information

Annex A - Excel file: Supplementary Table 1 - Sequences produced and used in the ENGAGE project. This table includes ENA submission numbers, isolates IDs and their use.

Annex A can be found in the online version of this output ('Supporting information' section): https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

Annex B - Excel file: Supplementary Table 2 - Phenotypic antimicrobial resistance (AMR) genes and resistance genes detected using SPAdes 3.9 vs Velvet 1.2 *de novo* assembly tools. This table includes the antimicrobial susceptibility results and AMR genes from 50 *Salmonella* isolates.

Annex B can be found in the online version of this output ('Supporting information' section: https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

Annex C - Excel file: Supplementary Table 3 - Detailed results for the ENGAGE benchmarking of genotypic *Salmonella* serotype prediction (general). This table includes list of serotypes, correlation, miscorrelation, no prediction, ambiguous results and summary graph.

Annex C can be found in the online version of this output ('Supporting information' section: https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

Annex D - Excel file: Supplementary Table 4: - Sequencing quality and detailed results for the ENGAGE benchmarking of genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6. This table includes sequencing quality, species, species-correlation, species-miscorrelation, species-no prediction, summary-species, serotype, serotype-correlation, serotype-miscorrelation, serotype-ambiguous, serotype-no prediction and summary serotype.

Annex D can be found in the online version of this output ('Supporting information' section: https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

Annex E - Excel file: Supplementary Table 5: - Sequence list and detailed results for the ENGAGE benchmarking of genotypic detection of antimicrobial (AMR) genes. This table includes sequence list from APHA, sequence list from DTU, comparison of AMR tools for *Salmonella* and *E. coli* dataset.

Annex E can be found in the online version of this output ('Supporting information' section: https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

Annex F - Excel file: Supplementary Table 6: - Sequence list and detailed results for the ENGAGE benchmarking of *Salmonella* Enteritidis phylogeny and *Campylobacter coli* phylogeny. This table includes 30 *Salmonella* isolates and 9 *Campylobacter* isolates.

Annex F can be found in the online version of this output ('Supporting information' section https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2018.EN-1431.

References

Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R, Underwood A, Jenkins C and Grant K, 2018. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. Bioinformatics, doi: 10.1093/bioinformatics/bty212.

Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, and Underwood A, 2016. MOST: a modified MLST typing tool based on short read sequencing. PeerJ 4, e2308. doi: 10.7717/peerj.2308.

Abbreviations

AMR	Antimicrobial resistance
APHA	Animal Plant Health Agency
BfR	Bundesinstitut für Risikobewertung
CDC	Centers for Disease Control and Prevention
CGE	Center for Genomic Epidemiology
CoC	Code of Conduct
DNA	Deoxyribonucleic acid
DTU	Technical University of Denmark, National Food Institute
eBG	e-Burst Groups
E. coli	*Escherichia coli*
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
ENA	European Nucleotide Archive
ENGAGE	Establishing Next Generation sequencing Ability for Genomic analysis in Europe
ESBL	Extended spectrum beta-lactamase
EU	European Union
EURL	European Union Reference Laboratory
EURL-AR	European Union Reference Laboratory for Antimicrobial Resistance
FBP	Foodborne pathogens
FDA	Food and Drug Administration
FWD	Food and Waterborne Disease
GMI	Global Microbial Identifier
IP	Intellectual Property
IZSLT	Istituto Zooprofilattico Sperimentale del Lazio e della Toscana M. Aleandri
IZSVE	Istituto Zooprofilattico Sperimentale delle Venezie
MCC	Matthew's Correlation Coefficient
MDR	Multidrug resistant
MoU	Memorandum of understanding
MTA	Material transfer agreement
NDA	Non-disclosure agreements
NGS	Next Generation Sequencing
NIPH-NIH	National Institute of Public Health – National Institute of Hygiene
NRL	National Reference Laboratory
NRL-AR	National Reference Laboratory for Antimicrobial Resistance
NVRI	National Veterinary Research Institute
PAHO	Pan American Health Organization
PHE	Public Health England
PT	Proficiency Testing
QC	Quality control
RIVM	The Netherlands National Institute for Public Health and the Environment
SNP	Single Nucleotide Polymorphism
SOP	Standard operating procedures
SPI	*Salmonella* pathogenicity Island
ST	Sequence Type
US	United States of America
USDA	US Department of Agriculture
VTEC	Verotoxin producing *E. coli*
WGS	Whole genome sequencing
WGST	Whole genome sequencing typing
WHO	World Health Organization
WP	Work Packages

Appendices

Appendix A - **Data collection**

Appendix B - **Guideline on how to get started**

Appeneix C - **SOPs for DNA extraction and library preparation (Illumina sequencing platform)**

Appendix D - **List of online bioinformatics tools and software used for capacity building**

Appendix E - **Benchmarking of *de novo* assembly tools: SPAdes 3.9 vs Velvet 1.2**

Appendix F - **Benchmarking of genotypic *Salmonella* serotype prediction (general)**

Appendix G - **Genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain – Method validation – Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures**

Appendix H - **Benchmarking of genotypic detection of antimicrobial resistance (AMR) genes**

Appendix I - **Benchmarking for *Salmonella* Enteritidis phylogeny**

Appendix J - **Benchmarking for *Campylobacter coli* phylogeny**

Appendix K - **Proof-of-concept projects**

Appendix L - **The ENGAGE Proficiency Test Report 2016**

Appendix M - **The ENGAGE Proficiency Test Report 2017**

Appendix A – Data collection

Based on the recorded phenotypic information combined with the existing epidemiological data and traditional molecular typing methods such as AMR and plasmid profiling, multi-locus variable-number tandem repeat analysis (MLVA), PFGE and multi locus sequence typing (MLST) based on 7 house-keeping genes, the consortium partners selected *Salmonella* spp. and *E. coli* strains for WGS. The overall aim was to retrospectively sequence a large number of strains in a comprehensive background strain collection to provide datasets for the benchmarking exercises, including strains for the specific pilot studies to provide high resolution information on the phylogenetic relationships amongst isolates from different EU countries and provide insight on pathogen evolution and acquisition and spread and AMR. The suggested criteria for strain selection were:

Salmonella spp.: *Salmonella* isolates with a specified AMR profile (e.g. ACSSuT, ASSuT, Cip/Nal, Cip without Nal, 3GCs, gentamicin) including isolates of human, food and animal origin isolated in the previous 10 years. To monitor the evolution, a few representative isolates of a specific serovar or AMR pattern older than 10 years were included in the selection. For animal isolates the selection was made on the basis of serovar per host/year base, to allow for proportional sampling of the number of isolates from specific animal host to the number of total isolates in that year and avoid overrepresentation, unless isolates are outbreak related. Strain selection for the *Salmonella* serotyping benchmarking exercise was based on inclusion of a collection of wide variation of serovars including commonly isolated serovars and rare serovars, seldom found.

E. coli: commensal *E. coli* and a few VTEC were included. The selections was done based on serotype and/or pulsotype, where available; *E. coli* isolates with a specific AMR profile including isolates of human, food and animal origin isolated in the previous 10 years were included in the selection. In addition, *E. coli* isolates from the 2015 EU surveillance in Germany, Denmark and the UK were sequenced.

All sequenced isolates (Table A.2) with the corresponding ENA submission numbers and project submission numbers are listed in the Supplementary Table 1 (Annex A). All sequences are thus publically available and also, most sequences have been included in ENGAGE proof-of-concept projects or ENGAGE benchmarking exercises as indicated in Annex A. Some sequences have not been included in ENGAGE projects but have been sequenced with the purpose of populating the ENA with sequences of relevant, interesting phenotypes.

DTU = 520

During the ENGAGE project period, DTU Food WGS typed a total of 520 genomes (212 *Salmonella* and 308 *E.coli*) isolated between 2008 and 2017. The sequenced *Salmonella* spp. isolates included serovars for inclusion in the proof of concept studies, i.e. 64 *S.* Derby strains covering the last five years and 5 *S.* Napoli sequences were included in the proof-of-concept studies conducted by IZSVe. Similarly, 81 *S.* Infantis strains were whole genome sequenced for the proof of concept study conducted by IZSLT. Due to all Danish *S.* Paratyphi B var. Java resistant to colistin being of German origin, we set up a proof of concept study with BfR to include the Danish genomes (17 *Salmonella* spp. sequences) in the project led by BfR. Forty-five of the sequenced *Salmonella* strains have not been included in ENGAGE projects but have been sequenced with the purpose of populating the ENA with sequences of relevant, interesting phenotypes. Of the 308 *E. coli* whole genome sequenced by DTU Food, 41 *E. coli* strains resistant to colistin were included in a DTU Food led project to characterize the strains and plasmids, 164 *E. coli* in the benchmarking exercise for genotypic detection of antimicrobial resistance (AMR) genes, 97 EBSL producing *E. coli* from the national AMR monitoring of animal species not included in the EU monitoring in the respective years (not included in an ENGAGE project but sequenced with the purpose of populating the ENA with sequences of relevant, interesting

phenotypes), and six *E. coli* ST-131 (not included in an ENGAGE project but sequenced with the purpose of populating the ENA with sequences of relevant, interesting phenotypes).

ENA Study ID: PRJEB23891, PRJEB23082, PRJEB18587, PRJEB14641, PRJEB18619, PRJEB22091, PRJEB14086

PHE = 500

As PHE routinely sequenced *Salmonella* since 2014, we selected 500 genomes of different *Salmonella* serovars to include in the serotyping benchmarking exercise to give a broad and deep overview of the *Salmonella* diversity seen between 2014 and 2015. To cover maximum diversity, we selected 5 strains for each serovar isolated during the period 2014-2015. All the 500 genomes selected were used in the serotyping benchmarking. As *Salmonella* Enteriditis and *Salmonella* Typhimurium are the most common serovars, more than 5 strains were selected from these two serovars to also capture the sequence type diversity. The final contribution included more than 100 different *Salmonella* serovars. The strain accession numbers are available on the SRA project PRJNA248792 (PHE global repository for *Salmonella* isolates). Table A.1 below summarizes the strain selection:

Table A.1: PHE strain selection

Serovars: each serovar mentioned contains # strains (see number in the column to the right)	Number of isolates per serovar
Typhimurium	16
Enteritidis	13
Paratyphi	8
Abony, Adelaide, Agama, Agbeni, Ago, Agona, Ajiobo, Alachua, Albany, Anatum, Bareilly, Blockley, Bovis-morbificans, Braenderup, Brandenburg, Bredeney, Cerro, Chester, Coeln, Colindale, Corvallis, Derby, Dublin, Durham, Eastbourne, Emek, Gaminara, Give, Gold-coast, Hadar, Haifa, Havana, Heidelberg, Hvittingfoss, Ibadan, Indiana, Infantis, Jangwani, Java, Javiana, Kedougou, Kentucky, Kenya, Kingston, Kottbus, Litchfield, Livingstone, London, Mbandaka, Mikawasima, Minnesota, Mississippi, Monschaui, Montevideo, Muenchen, Napoli, Newport, Nima, Ohio, Oranienburg, Oslo, Panama, Poona, Potsdam, Richmond, Rissen, Saint-paul, San-diego, Schwarzengrund, Senftenberg, Stanley, Stanleyville, Takoradi, Tel-el-kebir, Tennessee, Thompson, Typhi, Umbilo, Virchow, Wassenaar, Weltevreden	5 of each serovar
Aberdeen, Bispebjerg, Fluntern, Meleagridis, Muenster, Rubislaw, Vitkin	4
Altona, Bonn, Concord, Ealing, Khami, Kisangani, Manchester, Nottingham	3
Amager, Apapa, Glostrup	2

APHA = 439

APHA sequenced a total of 439 bacterial genomes including 339 *Salmonella* and 100 *E. coli* for inclusion in different ENGAGE studies. The selected isolates were: 94 "rare" *Salmonella* spp. isolates of different serovars isolated between 1988 and 1997, for inclusion in the *Salmonella* serotyping benchmarking exercise (Benchmarking exercise #2). The strains were selected from the APHA strain collection used for production of O and H anti-sera for the traditional *Salmonella* serotyping. Sixteen of the 94 "rare" *Salmonella* were not included in ENGAGE projects but were sequenced with the purpose of populating the ENA with sequences of relevant, interesting phenotypes. Further 123 *Salmonella* spp. isolates collected between 2006 and 2016 were selected for the AMR benchmarking exercise (Benchmarking exercise #4). In addition, 59 *S.* Dublin isolates collected from 2006-2016 were selected for inclusion in an EU-wide study lead by RIVM (external subproject lead by this affiliated partner, data not included in this report) 29 *S.* Derby isolates from 2006-2016 for inclusion in the

study lead by IZSVe and 34 *S.* Infantis strains from 2006-2016 for inclusion in the study lead by IZLST. *S.* Dublin, *S.* Derby, *S.* Infantis and isolates for the AMR benchmarking exercise were collected as part of the UK National Control Program (NCP) or as part of the passive surveillance conducted through the submission of isolates from clinical diagnostic samples. The 100 ESBL producing *E. coli* strains were part of the 2015 EU AMR monitoring programme for *E. coli* (not included in an ENGAGE project but sequenced with the purpose of populating the ENA with sequences of relevant, interesting phenotypes).

ENA study ID: PRJEB24311 (123 *Salmonella* spp isolates included in the AMR benchmarking study), PRJEB24308 (92 *Salmonella* spp. isolates included in the serotyping benchmarking), PRJEB23868 (2 *Salmonella* spp. isolates included in the serotyping benchmarking), PRJEB24107 (34 *S.* Infantis isolates included in the pilot study), PRJEB24103 (59 *S.* Dublin isolates included in the pilot study), PRJEB24097 (25 S. Derby isolates include in the pilot study), PRJEB24583 (4 S. Derby isolates included in the pilot study), and PRJEB21131 (100 *E. coli* isolates) (Supplementary Table 1 Annex A).

BfR = 382

In the two years of the ENGAGE project, BfR sequenced 382 *S. enterica* (290) and *E. coli* (92) isolates from animal, food and environmental sources using the *in-house* Illumina MiSeq and NextSeq sequencers. *Salmonella* isolates of the most common serovars as well as *Escherichia coli* isolates were selected based on conspicuous phenotypic antimicrobial resistance profiles. Altogether 16 *S.* Enteritidis, 49 *S.* Infantis, 35 *S.* Derby, 72 *S.* Typhimurium (monophasic and biphasic variants), 99 *S.* Paratyphi B variant Java, 1 *S.* Heidelberg, 1 *Salmonella* rough colony, 3 *S.* Newport, 3 *S.* Saintpaul, 11 *S.* Napoli and 92 *E. coli* were sequenced. The strains were uploaded to the ENA studies PRJEB23094 (*Salmonella*, 290 isolates) and PRJEB23572 (*Escherichia coli,* 92 isolates).

NIPH-NIH = 320

A total of 320 strains were sequenced from the National Institute of Public Health – NIH collection. The strain list included 31 *Escherichia coli* VTEC isolates collected between 2003 and 2017, representing most of the VTEC strains isolated in Poland in this time period. Of these, 24 were of the most prevalent serotype O157, 4 were O26 and 3 were not typed using classical phenotypic method (further performed analysis of WGS data showed that one of these strains was O157:H34, one O26:H11 and one H4 with no O type genes found).

Additionally 289 *Salmonella enterica* strains isolated between 2013 and 2017 were sequenced. Among them the most numerous group was monophasic *S.* Typhimurium (105 strains). Of other sequences *Salmonella enterica* 43 were *S.* Enteritidis, 18 *S.* Typhimurium, 16 *S.* Infantis, 4 *S.* Derby, 2 *S.* Paratyphi B and 1 *S.* Napoli (all included in the proof of concept partners projects). Additionally WGS was performed for 10 strains from other common serovars (*S.* Hadar, *S.* Virchow and *S.* Mbandaka) and 58 strains from rare serovars *S.* Schwarzengrund, *S.* Schleissheim, *S.* Oranienburg, *S.* Senftenberg, *S.* Bredeney, *S.* Poitiers and *S.* Vitkin. To check and point out the opportunities offered by WGS technology, 32 *S. enterica* strains, non-fully typed using classical phenotypic tests were sequenced. All sequences and metadata were uploaded on ENGAGE working space with no restrictions on their public availability.

ENA Study ID: PRJEB23743, PRJEB26541, PRJEB26514, PRJEB26516, PRJEB26511, PRJEB26513, PRJEB26517, PRJEB26518, PRJEB26520, PRJEB26519, PRJEB26504, PRJEB26523, PRJEB26528, PRJEB26529, PRJEB26506, PRJEB26530, PRJEB26527, PRJEB26510, PRJEB26503, PRJEB26515, PRJEB26521, PRJEB26507, PRJEB26505, PRJEB26539, PRJEB26522, PRJEB26524, PRJEB26525, PRJEB26526.

NVRI = 368

National Veterinary Research Institute has performed analyses of 368 bacterial isolates fulfilling predefined strain selection criteria. They were isolated between 2010 and 2017, but mostly (66%) in

2014-2016, when official AMR monitoring was launched. These included 182 *Salmonella* isolated along the food chain and 186 *E. coli* isolated from food (N = 52; from chicken, cattle and pig meat) and animal faeces (N = 134; isolates mostly from official AMR monitoring). At the very beginning of the project, the sequencing was outsourced (N = 202; Illumina MiSeq and HiSeq platform) whilst the rest of the strains were sequenced with Illumina MiSeq platform implemented at NVRI laboratories in 2017. All sequences and metadata were uploaded on ENGAGE working space with no restrictions on their public availability. The sequences are available for miniprojects run at NVRI and other partners within ENGAGE project. Specifically, NVRI projects focused on:

- identification of rare and atypical *Salmonella* serovars (N = 102)
- *ad hoc* project on WGS of *S.* Enteritidis related to egg-related outbreak (N = 13; confirmation of infection source)
- characterisation of colistin-resistant (*mcr-1* positive) *E. coli* isolated from animals in Poland (N = 80)

ENA study ID: PRJEB23993.

IZSLT = 382

IZSLT sequenced a total of 382 isolates of which 321 were *S. enterica* and 61 *E. coli:*

-Sequenced *E. coli* isolates, phenotypically colistin-resistant (N=55) or susceptible (N=6), were all included in the proof-of-concept project entitled "Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy, headed by IZSLT in collaboration with DTU.

-Sequenced *Salmonella* isolates consisted of 229 *S*. Infantis included in the proof-of-concept project entitled "***Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and antimicrobial resistance (AMR) genes",** headed by IZSLT in collaboration with DTU and 14 *Salmonella* of different serotypes included in the proof-of-concept project entitled "Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy, headed by IZSLT in collaboration with DTU. Seventy-eight *S.* Typhimurium isolates were also analysed for comparison purposes of detecting similarities and differences between colistin-resistant and colistin-susceptible isolates in these animal productions. All the *S.* Typhimurium isolates were collected in the context of Italian passive surveillance activities and active voluntary monitoring programmes conducted in meat-producing animals (2014-2015-2016), according to the sampling frame of the Dec. 2013/652/2015.

All sequenced data was uploaded to the ENGAGE working space and to ENA under the project accession numbers: PRJEB23778 and PRJEB23728

IZSVe = 449

IZSVe sequenced 449 *Salmonella* genomes, spanning years from 2005-2016 (older isolates for S. Napoli only, due to its low frequency of isolation), for different ENGAGE studies. The selected isolates were: 141 *S.* Napoli, 150 *S.* Derby, 88 monophasic *S.* Typhimurium, 30 *S.* Enteritidis and 40 other serovars (1 *S.* Abony, 1 *S.* Infantis, 2 *S.* Bredeney, 2 *S.* NA, 21 *S.* Stanleyville, 4 *S. enterica* subsp. Houtenae, 9 *S.* Kentucky, part of which were shared for the "rare serovar" project led by NVRI). All sequenced data were uploaded to ENA under project numbers PRJEB21283, PRJEB22761 (monophasic variant of *S.* Typhimurium), PRJEB23440 (*S*. Derby), PRJEB23407 (*S*. Napoli), PRJEB23485 (other serovars). The sequences were used for the projects run at IZSVE (*S*. Napoli, *S*. Derby, monophasic *S.* Typhimurium) and shared with other partners for their own projects.

Table A.2: Number of isolates sequenced

Institute	Number of strains in ENA*
DTU	520
PHE	500
APHA	439
BfR	382
NIPH-NIH	320
NVRI	368
IZLST	382
IZSVe	449
TOTAL	**3,360**

* Supplementary Table 1 (Annex A).

Abbreviations

A = Ampicillin
C= Chloramphenicol
S= Streptomycin
Su = Sulphonamide compounds
T = Tetracycline
Nal = Nalidixic acid
Cip = Ciprofloxacin
3GCs: 3rd generation cephalosporins

Appendix B – Guideline on how to get started

Considerations when designing a whole genome sequencing (WGS) service: From Sample to Result

Nucleic Acid Extraction

- Bacteria will require DNA extraction from isolates. Development of a robust protocol for nucleic acid extraction is critical but already available from many sources, e.g. ENGAGE (see Appendix C in this report). A key component to this is the extraction method. Many manual kits (e.g. Promega Wizard genomic DNA purification and Qiagen/Stratec genomic DNA purification spin columns) are suitable but it is critical to check that the resulting DNA is of sufficient quantity (Illumina recommendation 8-100 ng/µl, Illumina Nextera recommendation 1ng/µl then diluted to 0.2ng/µl). If more than a few tens of samples are expected to be processed per week a high throughput DNA purification system such as Qiasymphony, EZ1, SP/AS, Qiacube HT (Qiagen company) is recommended.
- Viruses present a greater problem for extraction and either an amplicon strategy (quick et al., 2016) or bait-based enrichment protocol (Depledge et al., 2011) is required.
- In both of these cases, work to assess yields from these protocols is essential so that a standard operating procedure (SOP) can be produced which, if followed, results in a high probability of the amount and quality of DNA being sufficient for subsequent library preparation.

Quantification

- Although it is possible that following a SOP generated from the previous step results in a consistent amount of DNA that does not require quantification, it is recommended that prior to library preparation, quantification is performed.
- Recommended instruments for quantification include the GloMax (high throughput) from Promega and Qubit (single tube) from ThermoFisher. The NanoDrop (ThermoFisher) is not recommended, due to lack of sufficient accuracy and consistency of the readings for the purposes of library preparation.

Library Preparation

- There are two main alternatives for library preparation:
    - Nextera – a kit from Illumina that makes the number of hands on steps minimal but has the disadvantages of giving slightly less uniform coverage compared to physical sheering methods (see below) and having a higher per sample cost. In addition, it is susceptible to being less efficient for genomes with a %GC content significantly different from 50%. Furthermore, Nextera is recommended for bacterial genomes but is probably not suitable for smaller viral genomes.
    - Physical sheering of DNA (e.g. from Covaris) and adaptor ligation. This is more technically challenging and the upfront cost is greater. However, the per sample cost is less and the uniformity of sequence coverage is better and less susceptible to variations in %GC.
- After preparing libraries for each sample including the addition of unique indices per sample, normalization of the quantity of DNA added per sample into the pooled tube (PAL) that will be

sequenced is essential in order to ensure each sample has adequate coverage. There are again at least two possible methods:

  - Using the Nextera XT kit Guide 150319425031942 following the protocol revision C (http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html) where a bead-based method ensures simple normalization of sample quantities that are added. Other bead based normalisation kits are available.
  - Measurement of the concentration of each sample.

- Ideally the fragment size of the libraries should also be measured before addition to the PAL tube in order to ensure the correct range for efficient sequencing (250 - 1000 bp). This can be achieved by fragment analysers such as LabChip from Perkin Elmer or BioAnalyser from Agilent.
- During library preparation a positive control comprising DNA from a known isolate (to check the effectiveness of library preparation and the absence of sample transposition) and ideally a negative control (to check for lack of contamination) should be included.

Sequencing

There are several short read sequencing technologies that are currently on the market, including Illumina and Ion Torrent™ Personal Genome Machine™ (from Thermo Fisher Scientific company) both of which have 'desktop' machines, the MiSeq and Ion Torrent™ Personal Genome Machine™ (PGM), respectively. These technologies also have larger capacity high throughput machines. When deciding which technology to use it is important to consider capacity and speed. Is turnaround time crucial and, if so, how much of a sequencing plate is required to be filled before it is sufficiently cost effective?

Whatever the technology used one critical step that should always be carried out and audited is the on-machine quality metrics calculation. It implies that the quality of the data as assessed by the machine is recorded as well as the quality of the final output fastq files. On the Illumina platform this will include metrics such as cluster density and percentage of clusters that pass/fail.

Post-sequencing data processing

- **Demultiplexing**
  When processing samples through the desktop machines, e.g. MiSeq, the processing of raw reads into per sample fastq files can occur on board the machine itself. However, for the higher throughput machines a server/computing infrastructure will likely be required.
- **Quality Control**
  Once the sequencing data has been demultiplexed it is critical that quality assessment is carried out (fastQC, see Appendix D, is recommended) and that, if necessary, poor quality data is removed using software such as Trimmomatic (Appendix D). At the very least adapter removal should be performed.
- **Analytical processes**
  Before embarking on the process of analysing samples to obtain results it is critical to think what the end point is and what result needs to be reported. Then a literature search can be carried out to assess how this can be best achieved. The list of software provided as part of the ENGAGE project (Appendix D) or the tools listed here (https://omictools.com/whole-genome-resequencing-category) will be a good place to start. A key consideration will be throughput. For any more than a few samples per week, a web based solution will probably not be suitable since it will be too person-hour expensive and difficult to audit and record. Alternatives for running analytic pipelines include:

- **Web services**
  A good example of this are the services at the Centre for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/all.php). These allow sample by sample processing and in some cases batch processing. However, tracking the result outcomes and version of software is challenging.
- **Galaxy**
  A wide range of tools as listed in this report (Appendix D) are available via the Galaxy website (https://usegalaxy.org/) and these can be chained together into pipelines. This offers a lot of flexibility although it is likely that downstream processing of the outputs will be required in order to make them ready for interpretation.
- **Infrastructure**
  If processing any more than a few samples a dedicated server running best practice software is desirable. However, this will require ongoing dedicated IT support and programmatic bioinformatics skills.

Whichever solution is chosen from the options listed above, the location for the long term storage of the data should be considered. Although data can be uploaded to the public nucleotide archives (e.g. EMBL-EBI (https://www.ebi.ac.uk/ena/submit/sra/#home) or NCBI https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/), it is likely that local storage of the files will be necessary. The amount of storage space required will be in the order of several terabytes. A resilient storage system recommended by local IT should be purchased unless they can give assurance of being able to store data of this magnitude.

Reporting

It is critical to think about the format and content of the final report that contains results derived from WGS. At an early stage consultation with the end-users (microbiologists, clinicians and epidemiologists) should be carried out in order to discuss what should be reported. The process by which the final outputs from the analytical pathways can be converted into a report should be planned at an early stage, to enable automatization.

Sample tracking and auditing

Throughout all these processes good record keeping and tracking of sample progress should be employed in order to allow construction of a full audit trail. A NGS sample LIMS (Laboratory Information and Management Systems) would be recommended such as the one listed here (https://omictools.com/lims-category).

References

Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P and Breuer J, 2011. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLOS ONE 6, e27805.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232. doi: 10.1038/nature16996

Companies' main websites

- QIAGEN: https://www.qiagen.com/gb/
- STRATEC: https://www.molecular.stratec.com/home
- THERMO FISHER: https://www.thermofisher.com
- PROMEGA: https://www.promega.co.uk/
- COVARIS: http://covaris.com/
- ILLUMINA: https://www.illumina.com/
- PERKIN ELMER: http://www.perkinelmer.com
- AGILENT: https://www.agilent.com/

Appendix C – SOPs for DNA extraction and library preparation when using the Illumina sequencing platform

Introduction

With rapid development of whole genome sequencing (WGS) analysis, many bacterial DNA extraction procedures have been tested. Here, we review three methods commonly used to extract bacterial DNA and library preparation methods for WGS.

DNA extraction methods

APHA uses an automated MagNA pure system (Roche Life Science) for routine DNA extraction. MagNA Pure LC 2.0 Instrument performs a majority of the extraction steps, including binding of DNA to magnetic glass particles, washing steps and elution of pure DNA. The purified DNA was analysed with respect to DNA integrity, recovery, purity and ability to amplify target sequence with LightCycler® 480 and LightCycler® (Roche Life Science) Carousel-Based Instruments. The product has since been withdrawn. Related products can be found the Roche website (*https://lifescience.roche.com/en_gb/products/magna-pure-24-instrument.html*).

In addition, APHA tested boilate method for bacterial DNA extraction for WGS. The boilate method developed for PCR templates (Queipo-Ortuno et al., 2008; Wimalarathna et al., 2013) has proven to be suitable for the preparation of libraries for WGS of *Mycobacterium bovis* at the APHA sequencing unit, however, testing by comparison to sequencing extracted DNA using MagNA Pure extraction method showed the boilates method not to be suitable for sequencing *Salmonella* spp. and *E. coli* genomes. The advantages of the boilate method include no requirements for special equipment or reagents, rapid preparation and a safe way to transport pathogenic isolates for sequencing and thus further method development outside this project will be carried out for potential use in *Salmonella* sequencing.

Since the start of the project, the MagNA Pure extraction system has been discontinued from production by Roche Life Sciences and therefore APHA adopted a similar magnetic separation protocol using KingFisher™ Duo Prime Magnetic Particle Processor (Thermofisher) (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/KingFisher_Duo_Prime_User_Manual_5400110.pdf).

At DTU, genomic DNA extraction is carried out using Easy-DNA™ Kit (Invitrogen, Thermofisher). The extraction method yields high-quality DNA with an average size between 100 kb and 200 kb, which is suitable for PCR, DNA hybridization, genomic DNA library construction, and other applications. The extraction procedure contains only 4 steps with no special equipment required.

(https://tools.thermofisher.com/content/sfs/brochures/713_021456_easydnapr_bro.pdf)

At PHE, genomic DNA extraction is carried out using an automated method to extract DNA from bacterial cells. The QIAsymphony DNA Investigator Kit (Qiagen) enables automated purification of genomic DNA from up to 96 samples from a wide range of starting material such as swabs, filters, casework, crime-scene samples and blood. Purification is fast and efficient, and purified DNA performs well in downstream analyses. This method requires QIAsymphony SP/AS instrument.

(https://www.qiagen.com/gb/resources/resourcedetail?id=b0c38b97-2200-4102-a2d5-ba99648fc9d5&lang=en)

Sequencing library preparation

DTU, APHA and PHE follow the Illumina NexteraXT library preparation manual (Illumina) (*https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteradna/nextera-dna-library-prep-reference-guide-15027987-01.pdf*).

Sequencing method

DTU, APHA and PHE use Illumina sequencing platforms including MiSeq, HiSeq and NextSeq.

Reference

Queipo-Ortuno MI, De Dios Colmenero J, Macias M, Bravo MJ and Morata P, 2008. Preparation of bacterial DNA template by boiling and effect of immunoglobulin G as an inhibitor in real-time PCR for serum samples from patients with brucellosis. *Clinical and Vaccine Iimmunology: CVI*, 15(2):293-296.

Wimalarathna HM, Richardson JF, Lawson AJ, Elson R, Meldrum R, Little CL, Maiden MC, McCarthy ND and Sheppard SK, 2013. Widespread acquisition of antimicrobial resistance among Campylobacter isolates from UK retail poultry and evidence for clonal expansion of resistant lineages. *BMC Microbiology*, 13:160.

APPENDIX C.1 STANDARD OPERATING PROCEDURE – DNA extraction (bacteria), DTU Food

Easy-DNA™ Kit (Invitrogen)

Prepared by: Rolf Sommer Kaas
Contact: rkmo@food.dtu.dk
Institution: DTU Food, Denmark
SOP-version: 1

Introduction

This protocol describes a bacterial DNA extraction method to extract DNA. Easy-DNA™ (Invitrogen) extraction method yields high-quality DNA with an average size between 100 kb and 200 kb, which is suitable for PCR, DNA hybridization, genomic DNA library construction, and other applications. In this method, protein and lipids are precipitated and extracted by the addition of solution B and chloroform. The solution is then centrifuged to separate it into two phases with a solid interface in between the phases. The DNA is in the upper. The extraction procedure contains only four steps with no special equipment required.

Sample Material

Bacterial culture

Equipment

- Heating block capable of maintaining 65°C
- Microcentrifuge
- Vortex Mixer

Reagents

- Easy-DNA™ Kit, cat.no. K1800-01 (Invitrogen)

Literature

Order info: https://www.fishersci.com/shop/products/kit-easy-dna/k180001

Procedure, DNA extraction (bacteria), DTU Food (Easy-DNA™ Kit (Invitrogen))

1. Resuspend a 10 µl loopful of bacterial cells streaked on blood agar in 200 µl PBS.
2. Add 350 µl of Solution A and incubate at 65°C for 10 min.
3. Add 350 µl of Solution B and vortex vigorously.
4. Add 500 µl of chloroform and vortex.
5. Spin at 20,000 g for 10 min at 4°C.
6. Transfer 300-500 µl of the upper phase into 1 ml of cooled ethanol.
7. Incubate on ice for 30 min.
8. Centrifuge at 20,000 g for 10-15 min at 4°C.
9. Wash with 500 µl of cooled 80% ethanol.
10. Centrifuge at maximum speed for 3-5 min at 4°C to remove ethanol.
11. Resuspend pellets in 100 µl of Tris-RNase (40 µg/ml) and incubate for 1 h at 37°C.
12. Store DNA samples at -20°C until required.

APPENDIX C.2 STANDARD OPERATING PROCEDURE – DNA extraction (bacteria), PHE

QiaSymphony DNA (Qiagen) extraction

Prepared by: Satheesh Nair, Craig Swift
Contact: Satheesh.Nair@phe.gov.uk; Craig.Swift@phe.gov.uk
Institution: PHE, UK
SOP-version: 1

Introduction
This protocol describes an automated method to extract DNA from bacterial cells. The QIAsymphony DNA Investigator Kit (Qiagen) enables automated purification of genomic DNA from 1–96 samples from a wide range of starting material, such as swabs, filters, casework or crime-scene samples, and blood on the QIAsymphony SP. Purification is fast and efficient, and purified DNA performs well in downstream analyses.

Sample Material
Bacterial culture

Equipment
- Heating block capable of maintaining 95°C
- Microcentrifuge tubes
- Vortex Mixer
- QIAsymphony SP/AS instrument

Reagents
- QIAsymphony DNA Investigator Kit

Literature
Application note:
https://www.qiagen.com/gb/resources/resourcedetail?id=b0c38b97-2200-4102-a2d5-ba99648fc9d5&lang=en

Procedure, DNA extraction (bacteria), PHE (QiaSymphony DNA extraction)
1. Transfer 700 µl of overnight culture into a Fortitude 96 well plate.
2. Spin at 3500 rpm for 20 min to collect bacterial cells.
3. Lyse cells with ATL buffer and Proteinase K.
4. Add 4 µl of RNase.
5. Heat inactivate for 95°C for 10 mins.
6. Transfer plate onto the QiaSymphony extractor.
7. Perform automated DNA extraction.
8. Store DNA samples at -20°C until required.

APPENDIX C.3 STANDARD OPERATING PROCEDURE – DNA extraction (bacteria), APHA

Bacterial DNA extraction with the MagNA Pure LC (Roche Life Science) system

Prepared by: Yue Tang
Contact: yue.tang@apha.gsi.gov.uk
Institution: APHA, UK
SOP-version: 1

Introduction

This protocol describes an automated method to extract DNA from bacterial cells. The isolation procedure is based on magnetic-bead technology. The samples are lysed by incubation with a special buffer containing chaotropic salts and Proteinase K. Magnetic Glass Particles are added and the DNA is bound to their surfaces. Unbound substances are removed by several washing steps, then the purified DNA is eluted. The MagNA Pure LC automatically performs all isolation and purification steps such as addition of Lysis/Binding buffer and magnetic glass particles (MGPs), binding of DNA to the MGPs, washing steps, elution of the pure DNA, and transfer to a cooled storage cartridge.

Sample Material

Bacterial culture

Equipment

- Heating block capable of maintaining 65°C
- Microcentrifuge tubes
- Vortex Mixer
- MagNA Pure LC 2.0 Instrument for 8-32 samples per run

Reagents

- MagNA Pure LC DNA Isolation Kit III (Bacteria, Fungi)

Literature

Application note:
MagNA Pure LC 2.0 Instrument has since been withdrawn. Related products can be found the Roche website *(https://lifescience.roche.com/en_gb/products/magna-pure-24-instrument.html*).

Procedure, DNA extraction (bacteria), APHA (bacterial DNA extraction with the MagNA Pure LC system)

1. Prepare 1.5 ml overnight cultures in LB broth from a single colony.
2. Spin to collect bacterial cells.
3. Wash cells with 500 µl of TE buffer.
4. Re-suspend cells in 100 µl of TE buffer.
5. Add 130 µl of Bacterial Lysis Buffer and 20 µl of Proteinase K.
6. Incubate at 65°C for 10 minutes.
7. Place 100 µl of sample mix in a sample cartridge.
8. Perform automated DNA extraction.
9. Store DNA samples at -20°C until required.

APPENDIX C.4 STANDARD OPERATING PROCEDURE – DNA extraction (bacteria), APHA

Preparation of cell boilates to extract DNA suitable for sequencing library preparation

Prepared by: Richard Ellis
Contact: richard.ellis@apha.gsi.gov.uk
Institution: APHA, UK
SOP-version: 1

Introduction
This protocol describes a rapid and inexpensive method to extract DNA from bacterial cells. This crude extract has proven to be suitable for the preparation of libraries for WGS of some bacteria such as Mycobacterium bovis.

Sample Material
Bacterial culture (single colony or pellet following centrifugation of broth culture)

Equipment
- Heating block capable of maintaining 95°C
- Microcentrifuge tubes
- Vortex Mixer

Reagents
- Molecular Biology Grade Water

General remarks
All bacterial cultures and boilates should be handled at the appropriate containment level. Once the inactivation of the bacteria by the heating process has been properly assessed and validated, boilates can be transferred to a lower containment level.

Literature
None

Procedure, DNA extraction (bacteria), APHA (preparation of cell boilates to extract DNA suitable for sequencing library preparation)
1. Dispense 100 µl of Molecular Biology Grade Water into Microcentrifuge tube.
2. Resuspend a single colony of bacteria (~3 mm^2) in the water.
3. Vortex for 15 s.
4. Briefly spin down to collect the liquid in the bottom of the tube.
5. Heat tube at 95°C for 10 minutes.
6. Spin down at 3500 rpm for 2 minutes to pellet cell debris.
7. Transfer supernatant to a fresh centrifuge tube (or well of a 96 well plate).
8. Store at -20°C until required.

APPENDIX C.5 STANDARD OPERATING PROCEDURE – Sequencing library preparation, APHA

Prepared by: Richard Ellis
Contact: Richard.Ellis@apha.gsi.gov.uk
Institution: APHA, UK
SOP-version: 1

Introduction

This protocol describes Nextera XT DNA Library Preparation (Illumina) with genomic DNA samples. The principle is that genomic DNA is randomly broken into small fragments (typically less than 1000 bp), before ligating sequencing primers to each end. Each of these ligated fragments are immobilized and clonally amplified, before denaturing. As complimentary bases are sequentially added to the single stranded template the sequence of nucleotides for each template is determined.

Sample Material

Genomic DNA samples

Equipment

- Heating block capable of maintaining 65°C
- 96-well microtiter plates
- Plate sealing film
- Centrifuge (capable of spinning 96-well plates between 100 x g and 1100 x g, room temperature)
- Thermocycler for 96-well plates
- Vortex mixer

Reagents

- Nextera XT library kit

Literature

Application note
http://emea.support.illumina.com/sequencing/sequencing_kits/nextera_xt_dna_kit/documentation.html

Procedure, Sequencing library preparation, APHA

1. Fragment DNA and then tags the DNA with adapter sequences in a single step.
2. Normalize gDNA.
3. Amplify libraries with 12 cycles of PCR.
4. Clean up libraries with AMPure XP beads.
5. Run 1 µl of undiluted library on an Agilent Technology 2100 Bioanalyzer to check libraries.
6. Normalize libraries to ensure equal representation.
7. Combine equal volumes of normalized libraries in a single tube for pooling libraries.

APPENDIX C.6 STANDARD OPERATING PROCEDURE – DNA sequencing using the MiSeq Instrument, APHA

Prepared by: Richard Ellis
Contact: Richard.Ellis@apha.gsi.gov.uk
Institution: APHA, UK
SOP-version: 1

Introduction

This procedure describes the steps required for the preparation of pooled Nextera® XT libraries for loading onto an Illumina® MiSeq Sequencing Platform. The Illumina MiSeq® system combines proven sequencing by synthesis (SBS) technology with a revolutionary workflow that enables you to go from DNA to analyzed data in as little as 8 hours. The MiSeq integrates cluster generation, sequencing, and data analysis on a single instrument.

Sample Material

Pooled libraries

Equipment

- The Illumina MiSeq desktop sequencer
- The Illumina Sequence Analysis Viewer software
- Vortex mixer

Reagents

- MiSeq Sequencing Kit v2 300 cycles (Illumina)

Literature

User guide:
http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-01.pdf

Procedure, DNA sequencing using the MiSeq Instrument, APHA

1. Mix 2.5 µl of diluted NaOH and 7.5 µl of a library pool for 5 min at RT.
2. Add 940 µl of hybridisation buffer to the denatured library and 50 µl PhiX Control.
3. Load 600 µl of the library mix onto the reagent cartridge.
4. Run the MiSeq system.
5. Wash the instrument with PR2 buffer.
6. Inspect cluster density for the data output quality.

Appendix D – List of online bioinformatics tools and software used for capacity building (status January 2018)

This document describes the most commonly used software and algorithms for processing whole genome sequencing. It is divided into categories, which describe the key processes for analysing short read data. Tools of particular interest will be tag with a specific character (historical†, commonly used*, easy to run#, etc). We are aware that the list is not complete, and that we present the status as of January 2018. It should be also taken into account that the area is continuously under development and new tools, not included here, will be released.

Most of the presented tools are command line based. In order to use them, you will need to install them on your infrastructure. We highly recommend that you ensure to have proper settings for your infrastructure (i.e. storage capacity and memory to run tools/software) as some of them require a lot of resources. In case you want to try these softwares and do not have infrastructure, we can recommend you to run Bio-Linux using a Virtual Box.[9]

Quality Assessment and Trimming

This is the process by which the quality of fastq files is determined and subsequent optional trimming of the data to trim or remove poor quality reads is carried out.

- Trimmomatic*
  - http://www.usadellab.org/cms/index.php?page=trimmomatic
  - Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bolger, A. M., Lohse, M., & Usadel, B. (2014). *Bioinformatics*, btu170.
  - Windows, Mac OS X and Linux
  - A flexible read trimming tool that will remove Illumina adapters, reads below a certain length and low quality ends of the read
  - *Comments*: Trimming occurs in the order, which the steps are specified on the command line. It is recommended in most cases that adapter clipping, if required, is done as early as possible. Options will strongly depend on the data you used i.e. single end, paired end.
- FastQC*#
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - Windows, Mac OS X and Linux
  - A quality control tool for assessing the quality of NGS data
  - *Comments*: tool available both online/command line. If running more than few samples using the command line is recommended. Interpretation of the results is linked to the sequencing method used. The online documentation details all the warnings and how to interpret them.
- Seqtk
  - https://github.com/lh3/seqtk
  - Windows, Mac OS X and Linux
  - Tool for processing sequences in the FASTA or FASTQ format that can be used for adapter removal and trimming of low-quality bases
- FastX
  - http://hannonlab.cshl.edu/fastx_toolkit/

[9] Installation of Virtual Box tutorial is included at the end of this Appendix.

- Windows, Mac OS X and Linux
- Toolkit for FASTQ and FASTA pre-processing that can be used for trimming, clipping, barcode splitting, formatting and quality trimming.

Assembly

This is the process of joining short/long reads into longer contigs (contiguous lengths of DNA) without the need for a reference sequence.

- VelvetK
  - http://www.vicbioinformatics.com/software.velvetk.shtml
  - Windows, Mac OS X and Linux
  - Perl script to estimate best k-mer size to use for your Velvet de novo assembly.
- VelvetOptimiser
  - http://www.vicbioinformatics.com/software.velvetk.shtml
  - Mac OS X and Linux
  - Perl script to assist with optimising the assembly.
  - *Comments*: optimisation can be made using different metrics (e.g. with best N50, best coverage…)
- KmerGenie
  - http://kmergenie.bx.psu.edu/
  - Informed and Automated k-Mer Size Selection for Genome Assembly. Chikhi R., Medvedev P. HiTSeq 2013.
  - Windows, Mac OS X and Linux
  - Best k-mer length estimator for single-k genome assemblers like velvet.
- Khmer
  - http://khmer.readthedocs.io/en/v2.0/
  - The khmer software package: enabling efficient nucleotide sequence analysis. Crusoe et al., 2015. F1000 http://dx.doi.org/10.12688/f1000research.6924.1
  - Linux and Mac OS X
  - Set of command-line tools for dealing with large and noisy datasets to normalise and scale the data for more efficient genome assembly.
- Minia
  - http://minia.genouest.org/
  - Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Chikhi, Rayan and Rizk, Guillaume. Algorithms for Molecular Biology, BioMed Central, 2013, 8 (1), pp.22.
  - Windows, Mac OS X and Linux
  - Short-read assembler based on a de Bruijn graph for low-memory assembly.
- SPAdes*#
  - http://cab.spbu.ru/software/spades/
  - SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, Anton Bankevich, Sergey Nurk, Dmitry Anipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Journal of Computational Biology 19(5) (2012), 455-477. doi:10.1089/cmb.2012.0021
  - Mac OS X and Linux
  - Short and hybrid-long read assembler based on a de Bruijn graph that also performs error correction and is a multi-k genome assembler.
  - *Comments*: Illumina Paired reads (2*150 and 2*250) need to be assemble with the specific option --careful (see application note for full details) to get the best assembly possible

- Velvet†*
  - https://www.ebi.ac.uk/~zerbino/velvet/
  - Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Daniel R. Zerbino and Ewan Birney. Genome Res. May 2008 18: 821-829; Published in Advance March 18, 2008, doi:10.1101/gr.074492.107
  - Linux
  - De novo short read genome assembler with error correction to produce high quality unique contigs.
  - *Comments*: parameters can be difficult to select, some scripts have been developed and are working well to help choose the best parameters. Optimisation of the option should be used: VelvetOptimiser or VelvetK
- Canu
  - http://canu.readthedocs.io/en/stable/index.html
  - Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Adam M. Phillippy doi: http://dx.doi.org/10.1101/071282
  - Windows, Mac OS X and Linux
  - Long-read assembler designed for high-noise data such as that generated by PacBio or Oxford Nanopore MinION. Canu also performs error correction.
  - *Comments*: specifically designed to work with long read
- Unicycler
  - https://github.com/rrwick/Unicycler
  - Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, Kathryn E. Holt , Published in PLoS Comput Biol (2017) https://doi.org/10.1371/journal.pcbi.1005595
  - Mac OS X and Linux
  - Unicycler is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser. It can also assemble long-read-only sets (PacBio or Nanopore) where it runs a miniasm+Racon pipeline. For the best possible assemblies, give it both Illumina reads and long reads, and it will conduct a hybrid assembly.
  - *Comments*: use mainly as hybrid assembly for long read associated with Illumina read. Well documented with a Wiki-tutorial https://github.com/rrwick/Unicycler/wiki/Tips-for-finishing-genomes
- Bandage#
  - http://rrwick.github.io/Bandage/
  - Bandage: interactive visualization of de novo genome assemblies. Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bioinformatics (2015) 31 (20): 3350-3352 first published online June 22, 2015 doi:10.1093/bioinformatics/btv383
  - Linux and Mac
  - Program for visualising de novo assembly graphs by displaying connection which are not present in the contigs file for assembly assessment.
  - *Comments*: possibility to use blast inside the software to annotate regions of interest. Can help determine relations between contigs.

Annotation

The process which takes the raw sequence of contigs resulting from assembly and marks it with features such as gene names and putative functions.

- Prokka*#
  - http://www.vicbioinformatics.com/software.prokka.shtml
  - Prokka: rapid prokaryotic genome annotation. Seemann T. Bioinformatics. 2014 Jul 15;30(14):2068-9. PMID:24642063

- Windows, Mac OS X and Linux
- Software tool for the rapid annotation of prokaryotic genomes.

- RAST
  - http://rast.nmpdr.org/
  - The RAST Server: Rapid Annotations using Subsystems Technology. Aziz RK et al.. BMC Genomics, 2008
  - Online tool
  - Fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes.
- Genix
  - http://labbioinfo.ufpel.edu.br/cgi-bin/genix_index.py
  - Online tool
  - Fully automated pipeline for bacterial genome annotation.
- Prodigal
  - https://github.com/hyattpd/Prodigal/wiki/Introduction
  - Hyatt, Doug et al. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." BMC Bioinformatics 11 (2010): 119. PMC. Web. 25 Apr. 2018.
  - Windows, Mac OS X, GenericUnix (Linux)
  - Prodigal is a software is a protein-coding gene prediction software tool for bacterial and archaeal genomes
- NCBI Prokaryotic Genome Annotation Pipeline (PGAP)
  - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
  - Online tool – available for GenBank submitters only
  - PGAP is a pipeline for prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements

Alignment or sequence searching

Tools to align a sequence to other sequences locally or against publically available nucleotide or protein archives.

- BLAST[†#*]
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Basic local alignment search tool. Stephen F. Altschul,Warren Gish,Webb Miller,Eugene W. Myers,David J. Lipman. Journal of Molecular Biology, Volume 215, Issue 3, 5 October 1990, Pages 403-410
  - Windows, Mac OS X and Linux
  - Search tool to find regions of similarity between biological sequences through alignment and calculating statistical significance.
  - Comments: classic methods to search for specific sequence. Different version can be used such as blastn or megablast depending on the similarity between biological sequences. Possibility to create local specific database with makeblastdb.
- MUMmer
  - http://mummer.sourceforge.net/
  - Versatile and open software for comparing large genomes. A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg, Nucleic Acids Research (2002), Vol. 30, No. 11 2478-2483.
  - Windows, Mac OS X and Linux
  - A system for rapidly aligning entire genomes and finding matches in DNA sequences.
- Clustal suite – ClustalO and ClustalW
  - http://www.clustal.org

- Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22, 4673-4680.
- Sievers et al. (2011) Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. Molecular Systems Biology, 10.1038/msb.2011.75
- Windows, Mac OS X and Linux and online (webservers)
- Software that preforms sequences alignments. Mostly based on sequence weighting, position-specific gap penalties and weight matrix choice.
- Comments: ClustalO is usually present as performing better (faster and more accurate) than the original version of ClustalW.

- MUSCLE*#†
  - https://www.drive5.com/muscle/
  - Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, (5) 113
  - Windows, Mac OS X and Linux and online (webservers)
  - Software for multiple alignment of protein sequences.

Mapping

Alignment of short reads against a reference sequence so that amount of coverage or variations compared to the reference can be assessed.

- BWA*#
  - http://bio-bwa.sourceforge.net/
  - Fast and accurate short read alignment with Burrows-Wheeler Transform. Li H. and Durbin R. (2009) Bioinformatics, 25:1754-60. [PMID: 19451168]
  - Windows, Mac OS X and Linux
  - Software package for mapping low-divergent sequences against a large reference genome using the Burrows-Wheeler transform algorithm.
- Bowtie 2*#
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
  - Fast gapped-read alignment with Bowtie 2. Langmead B, Salzberg S. Nature Methods. 2012, 9:357-359.
  - Windows, Mac OS X and Linux
  - Tool for aligning sequencing reads to long reference genomes also based on the Burrows-Wheeler transform algorithm.
- Tablet
  - https://ics.hutton.ac.uk/tablet/
  - Using Tablet for visual exploration of second-generation sequencing data. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. 2013. Briefings in Bioinformatics 14(2), 193-202.
  - Windows, Mac OS X and Linux
  - Comments: Lighweight, high-performance graphical viewer for next generation sequence assemblies and alignments that can be used to view mapping.

Assembly refinement

Process of curating assembly by re-using reads and re-mapping steps

- Pilon
  - https://github.com/broadinstitute/pilon/wiki
  - Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant

Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. doi:10.1371/journal.pone.0112963
    - Windows, Mac OS X, Linux
    - Java based software that automatically improve draft assemblies. Find variation among strains, including large event detection.
    - Comments: assembly need to be performs prior to use the software.
- FGAP
    - https://github.com/pirovc/fgap
    - Piro, Vitor C et al. "FGAP: An Automated Gap Closing Tool." BMC Research Notes 7 (2014): 371. PMC
    - Online servers or Linux and Mac OS X
    - FGAP is a tool for closing gaps of draft genome. It uses BLAST to align multiple contigs against a draft genome assembly aiming to find sequences that overlap gaps. The algorithm selects the best sequence to fill and eliminate the gaps.

Assembly statistics and quality assessment

- Quast*#
    - http://quast.sourceforge.net/
    - Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086 First published online: February 19, 2013
    - Linux, MAC OS X and online servers
    - QUAST is a tool design to evaluate assembly. Calculates metrics such as N50, number of contigs, length of assemblies, GC content.
    - Comments: this tool accepts multiple assemblies and is suitable for comparing assemblies.

Variant Calling

Variant calling is the process by which variants (differences) are identify from sequence data. It usually follows the step of mapping reads against a reference.

- SAMtools*
    - http://samtools.sourceforge.net/
    - The Sequence alignment/map (SAM) format and SAMtools. Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) Bioinformatics, 25, 2078-9. [PMID: 19505943]
    - Windows, Mac OS X and Linux
    - Toolkit that provides various utilities for manipulating alignments in the SAM format and also can be used for generating consensus sequences and variant calling
- GATK*
    - https://software.broadinstitute.org/gatk/
    - The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. Genome Res. September 2010 20: 1297-1303; Published in Advance July 19, 2010, doi:10.1101/gr.107524.110
    - Windows, Mac OS X and Linux
    - Toolkit with a primary focus on variant discovery and genotyping.

- Picard
  - http://broadinstitute.github.io/picard/
  - Windows, Mac OS X and Linux
  - A set of command line tools (in Java) for manipulating high-throughput sequencing data and formats.
  - Comments: command line only, but helpful to convert/sort and use different output bam, sam…
- Varscan (version 2)
  - http://dkoboldt.github.io/varscan/
  - VarScan 2: Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing Genome Research DOI: 10.1101/gr.129684.111
  - Windows, Linux and Mac OS X
  - A set of command line tools running with Java that detects different kind of variants such as Germline variants (SNPs an dindels), Multi-sample variants (shared or private) in multi-sample datasets (with mpileup), Somatic mutations, Somatic copy number alterations (CNAs).

Phylogenetic analysis

Assessment of the evolutionary relationship between strains using either distance-based or Bayesian methodologies

- RaxML*
  - http://sco.h-its.org/exelixis/web/software/raxml/index.html
  - RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. A. Stamatakis. Bioinformatics (2014) 30 (9): 1312-1313.
  - Windows, Mac OS X and Linux
  - Randomized Axelerated Maximum Likelihood program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees.
  - Comments: maximum-likelihood methods give more resolution/accuracy than FastTree but take longer to run. Substitution models can be use as parameters.
- FastTree*#
  - http://www.microbesonline.org/fasttree/
  - FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). Molecular Biology and Evolution 26:1641-1650, doi:10.1093/molbev/msp077.
  - Windows, Mac OS X and Linux
  - Comments: Faster tool for speedy inference of approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. Particularly useful to quickly generate trees.
- CSI Phylogeny*#
  - https://cge.cbs.dtu.dk/services/CSIPhylogeny/
  - Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. Rolf S. Kaas , Pimlapas Leekitcharoenphon, Frank M. Aarestrup, Ole Lund. PLoS ONE 2014; 9(8): e104984.
  - Comments: Online tool, easy to use and configure. Tool to call SNPs, filter the SNPs and to do site validation and inference of phylogeny through a graphical user interface.
- Harvest
  - https://www.cbcb.umd.edu/software/harvest

- The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Treangen TJ, Ondov BD, Koren S, Phillippy AM. Genome Biology, 15 (11), 1-15
- Windows, Mac OS X and Linux
- Suite of core-genome alignment and visualization tools for quickly analysing thousands of intraspecific microbial genomes, including variant calls, recombination detection, and phylogenetic trees.
- Comments: parsnp from this tool can compute trees based on very large number of assembled genomes.

- Gubbins
  - http://sanger-pathogens.github.io/gubbins/
  - Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Croucher N. J., Page A. J., Connor T. R., Delaney A. J., Keane J. A., Bentley S. D., Parkhill J., Harris S.R. doi:10.1093/nar/gku1196, Nucleic Acids Research, 2014.
  - Windows, Mac OS X and Linux
  - Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) is an algorithm that iteratively identifies loci containing elevated densities of base substitutions while concurrently constructing a phylogeny based on the putative point mutations outside of these regions.
  - Comments: detection of recombination and generation of phylogeny. Depending on the number of genomes to analyse, this tool can be really long to run.
- BEAST
  - http://beast.bio.ed.ac.uk/
  - Bayesian phylogenetics with BEAUti and the BEAST 1.7. Drummond AJ, Suchard MA, Xie D & Rambaut A (2012) Molecular Biology And Evolution 29: 1969-1973.
  - Windows, Mac OS X and Linux
  - Cross-platform program for Bayesian analysis of molecular sequences using MCMC.
  - Comments: can be use to generate phylogeny based on prior information like time. Useful if you expect some time-relation in your phylogeny but really long to run.
- FigTree*#
  - http://tree.bio.ed.ac.uk/software/figtree/
  - Windows, Mac OS X and Linux
  - A graphical viewer of phylogenetic trees and program for producing publication-ready figures of trees.
  - Comments: easy tools to visualise/manipulate trees
- I-TOL*#
  - https://itol.embl.de/
  - Letunic I and Bork P (2016) Nucleic Acids Res doi: 10.1093/nar/gkw290 Interactive Tree Of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees
  - Online server
  - I-TOL Interactive Tree Of Life is an online tool for the display, annotation and management of phylogenetic trees.
  - Comments: This is only visualisation. Registration to have a workspace to save/manipulate tree. Really powerful to view large/complex tree. An extensive range of annotation available.
- Mega†
  - http://www.megasoftware.net/
  - MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Kumar S, Stecher G, and Tamura K (2016) Molecular Biology and Evolution 33:1870-1874
  - Windows, Mac OS X and Linux

- Comments: Sophisticated and user-friendly software suite for analysing DNA and protein sequence data from species and populations. Contains building tree algorithms.

Virulence and antimicrobial resistance gene prediction

Inference of potential for a virulent phenotype or resistance to an antimicrobial based on nucleotide sequences.

Virulence prediction

- PathogenFinder
  - https://cge.cbs.dtu.dk/services/PathogenFinder/
  - PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O. (2013) PLoS ONE 8(10): e77302.
  - Online tool
  - Web-server for the prediction of bacterial pathogenicity by analysing the input proteome, genome, or raw reads provided by the user.

Antimicrobial resistance prediction

- Resfinder
  - https://cge.cbs.dtu.dk//services/ResFinder/
  - Identification of acquired antimicrobial resistance genes. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. J Antimicrob Chemother. 2012 Jul 10
  - Online tool
  - Web-server that identifies acquired antimicrobial resistance genes in total or partial sequenced isolates of bacteria.
- ARIBA
  - https://github.com/sanger-pathogens/ariba
  - ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, Simon R Harris. doi: https://doi.org/10.1099/mgen.0.000131
  - ARIBA (Antimicrobial Resistance Identification By Assembly), identifies AMR-associated genes and single nucleotide polymorphisms directly from short reads
  - Comments: can also be used for MLST calling, you need to provide your reference set.
- KmerResistance
  - https://cge.cbs.dtu.dk/services/KmerResistance-2.2/
  - Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data
    Philip T.L.C. Clausen, Ea Zankari, Frank M. Aarestrup, Ole Lund
    Journal of Antimicrobial Chemotherapy. 2016
  - KmerResistance is a tool on a web-server that identifies antimicrobial resistance genes based on read mapping. It examines the co-occurrence of k-mers between the WGS data and a database of resistance genes.
  - Comments: reads mapping based detection of AMR genes is a great alternative to assembly based methods.
- SRST2
  - https://github.com/katholt/srst2
  - SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Inouye et al. Genome Medicine. 2014
  - Linux, Mac OS X

        - Short Read Sequence Typing for Bacterial Pathogens (SRST2) is designed to take Illumina sequence data, a MLST database and/or a database of gene sequences (e.g. resistance genes, virulence genes, etc) and report the presence of STs and/or reference genes.
    - GeneFinder
        - In-house tool developed by Public Health England (PHE, UK)
        - Genefinder software is a tool to determine presence and absence of genes and retrieve specific sequence variations from NGS paired-end fastq files, using a set of reference sequences in FASTA format
    - CARD: The Comprehensive Antibiotic Resistance Database (not a tool)
        - https://card.mcmaster.ca/home
        - CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Jia et al. Nucleic Acids Res. 2017. DOI:10.1093/nar/gkw1004
        - Contain an online pipeline RGI (Resistance Gene Identifier) to identify/query the CARD database for your genomes.
        - Database of resistance genes, their products and associated phenotypes.
        - Comments: useful resource for AMR. RGI need assemblies to run.

Species and serovar identification

Tools and software that uses various algorithms methods to identify a species by using reads or assembly and predict serovar. These software relies on databases to predict species or serovar.

- Kraken
    - https://ccb.jhu.edu/software/kraken/
    - Kraken: ultrafast metagenomic sequence classification using exact alignments. Wood DE, Salzberg SL. Genome Biology 2014, 15:R46.
    - Linux
    - System for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomics studies.
- MetaPhlan2
    - https://bitbucket.org/biobakery/metaphlan2
    - MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower & Nicola Segata. Nature Methods 12, 902-903 (2015)
    - Linux – command line
    - MetaPhlAn 2: Metagenomic Phylogenetic Analysis - profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species-level. The StrainPhlAn module allows to perform accurate strain-level microbial profiling.
- Kmerfinder
    - https://cge.cbs.dtu.dk/services/KmerFinder/
    - Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. Hasman H, Saputra D, Sicheritz-Pontén T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. J Clin Microbiol. 2014 Jan;52(1):139-46.
    - Online tools and standalone Linux version
    - Tool to identify species from an assembly or reads based on k-mer detection, searching k-mer from a pre-build database (bacterial, fungi viruses…).
    - Comments: the online tool can be used with different database
- SISTR[#]
    - https://lfz.corefacility.ca/sistr-app/
    - The Salmonella In Silico Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. Catherine Yoshida, Peter Kruczkiewicz, Chad R. Laing, Erika J. Lingohr, Victor P.J. Gannon, John H.E. Nash, Eduardo N. Taboada. PLoS ONE 11(1): e0147101. doi: 10.1371/journal.pone.0147101

- Web based application or standalone version on Linux and Mac OS X
- SISTR is a prediction software that predict serovar predictions from whole-genome sequence assemblies by determination of antigen gene. It also includes MLST, rMLST and cgMLST gene alleles prediction.

- SeqSero
  - http://www.denglab.info/SeqSero
  - Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015 May;53(5):1685-92.PMID:25762776
  - Online webserver and command line (Unix based)
  - SeqSero is a pipeline for *Salmonella* serotype determination from raw sequencing reads or genome assemblies
- MOST
  - https://github.com/phe-bioinformatics/MOST
  - Tewolde, Rediat et al. "MOST: A Modified MLST Typing Tool Based on Short Read Sequencing." Ed. Nicholas Loman. PeerJ 4 (2016): e2308. PMC. Web. 25 Apr. 2018.
  - Command line (Unix based)
  - MOST is a software derived from SISTR that assign MLST profile and infer Salmonella serotyping from bacterial genomic short read sequence data
  - Comments: require MLST database, detects novel allele if not present in the database, quality of the results assess by different metrics. Can be run in a Galaxy environment.
- Serotypefinder
  - https://cge.cbs.dtu.dk/services/SerotypeFinder/
  - Joensen, K. G., A. M. Tetzschner, A. Iguchi, F. M. Aarestrup, and F. Scheutz. 2015. Rapid and easy in silico serotyping of Escherichia coli using whole genome sequencing (WGS) data. J.Clin.Microbiol. 53(8):2410-2426. doi:JCM.00008-15 [pii];10.1128/JCM.00008-15
  - Online tool
  - SerotypeFinder identifies the serotype in total or partial sequenced isolates of E. coli.

Comparative genomic tools

Comparison of multiple genomes to determine regions of similarity or difference either on a gene-by gene basis or across the whole genome.

- BEDTools
  - http://bedtools.readthedocs.io/en/latest/index.html
  - BEDTools: a flexible suite of utilities for comparing genomic features. Aaron R. Quinlan and Ira M. Hall. Bioinformatics (2010) 26 (6): 841-842 first published online January 28, 2010 doi:10.1093/bioinformatics/btq033
  - Mac OS X and Linux
  - Toolkit for the manipulation of genome data for genomic analysis tasks on genomic intervals from multiple files.
- Roary
  - https://sanger-pathogens.github.io/Roary/
  - Roary: Rapid large-scale prokaryote pan genome analysis. Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill. Bioinformatics, 2015;31(22):3691-3693 doi:10.1093/bioinformatics/btv421.
  - Windows, Mac OS X and Linux
  - High speed stand-alone pan genome pipeline, which takes annotated assemblies in GFF3 format and calculates the pan genome.

- Mauve
  - http://darlinglab.org/mauve/mauve.html
  - Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Aaron C.E. Darling, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. Genome Res. July 2004 14: 1394-1403; doi:10.1101/gr.2289704
  - Windows, Mac OS X and Linux
  - Interactive genome alignment software that allows for easy browsing of multiple genomes to look for similarities and differences.
- ACT
  - http://www.sanger.ac.uk/science/tools/artemis-comparison-tool-act
  - ACT: the Artemis Comparison Tool. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG and Parkhill. Bioinformatics (Oxford, England) 2005;21;16;3422-3. PUBMED: 15976072; DOI: 10.1093/bioinformatics/bti553
  - UNIX, MacOS and Windows
  - Java application for displaying pairwise comparisons between two or more DNA sequences and allowing browsing of detailed annotation
- BRIG
  - http://brig.sourceforge.net/
  - BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. NF Alikhan, NK Petty, NL Ben Zakour, SA Beatson (2011). BMC Genomics, 12:402. PMID: 21824423
  - UNIX, MacOS and Windows
  - Image generating software that displays circular blast comparisons between a large number of genomes or DNA sequences

- EasyFig
  - http://mjsull.github.io/Easyfig/
  - Easyfig: a genome comparison visualiser. Sullivan MJ, Petty NK, Beatson SA. (2011) Bioinformatics; 27 (7): 1009-1010.PMID: 21278367
  - UNIX, MacOS and Windows
  - Python application for creating linear comparison figures of multiple genomic loci with an easy-to-use graphical user interface (GUI)
- SeqFindR
  - https://github.com/mscook/SeqFindR
  - UNIX and MacOS
  - Tool to easily create informative genomic feature plots by detecting the presence or absence of genomic features from a database in a set of genomes

Cloud Services

If infrastructure is not available the cloud based services are worth considering

- **Genomics-Specific**
- MRC CLIMB
  - http://www.climb.ac.uk/
  - Microbial bioinformatics cyber-infrastructure.
- Genomics Virtual Laboratory
  - https://www.gvl.org.au/
  - A genomics-specific version of Galaxy
- Galaxy
  - https://usegalaxy.org/
  - an open source, web-based platform for data intensive biomedical research.

- **Non-Genomics Specific**
- Amazon Web Services

- https://aws.amazon.com
- Pay per usage cloud computing managed by amazon.com for temporary computing of big data

- Azure (Microsoft)
  - https://azure.microsoft.com/en-us/
  - Multiple services divided into the following categories: AI + Machine Learning, Analytics, Compute, Containers, Databases, Developer Tools, DevOps, Identity, Integration, Internet of Things, Management Tools, Media, Migration, Mobile, Networking, Security, Storage, Web

Commercial software

- Bionumerics Seven
  - http://www.applied-maths.com/applications
  - Offers a range of tools to analyse sequence data including MLST, wgMLST, AMR profiling, wgSNPs.
- Ridom SeqSpehere +
  - http://www.ridom.de/seqsphere/index.shtml
  - Software design to analyse NGS data by using MLST/cgMLST

Blogs and Twitter

A lot of useful information in the rapidly evolving field of bioinformatics can be gained by following bioinformaticians on twitter or reading their blogs.

- Blogs
  - Bits and bugs https://bitsandbugs.org/
  - Loman Labs http://lab.loman.net/page3/
  - Opinionomics http://www.opiniomics.org/
  - The genome factory http://thegenomefactory.blogspot.co.uk/
  - Simpson Lab Blog http://simpsonlab.github.io/2016/08/23/R9/
  - Jonathon Eisen's Lab https://phylogenomics.wordpress.com/
  - Living in an Ivory Basement http://ivory.idyll.org/blog/
  - Holt Lab https://holtlab.net/
  - Heng Li's blog https://lh3.github.io/
  - The Darling lab http://darlinglab.org/blog/
  - The Quinlan Lab http://quinlanlab.org/

- Help pages
  - https://www.biostars.org
  - http://stackoverflow.com/

- Bioinformaticians to follow on Twitter
  - @pathogenomenick @BioMickWatson @flashton2003 @WvSchaik @mattloose @torstenseemann @tomrconnor @MikeyJ @jaredtsimpson @aphillipy @BillHanage @happy_khan @daanensen @jennifergardy @genomiss @Becctococcus @phylogenomics @ctitusbrown @DrKatHolt @ZaminIqbal @TimDallman @bioinformant @LaurenCowley4 @gkapatai @keithajolley @froggleston @lexnederbragt @jacarrico @biocomputerist @mjpallen @Bio_mscook @bawee @lh3lh3 @andrewjpage @aaronquinlan @koadman @Maxi_Zu

Installation of Virtual Box tutorial

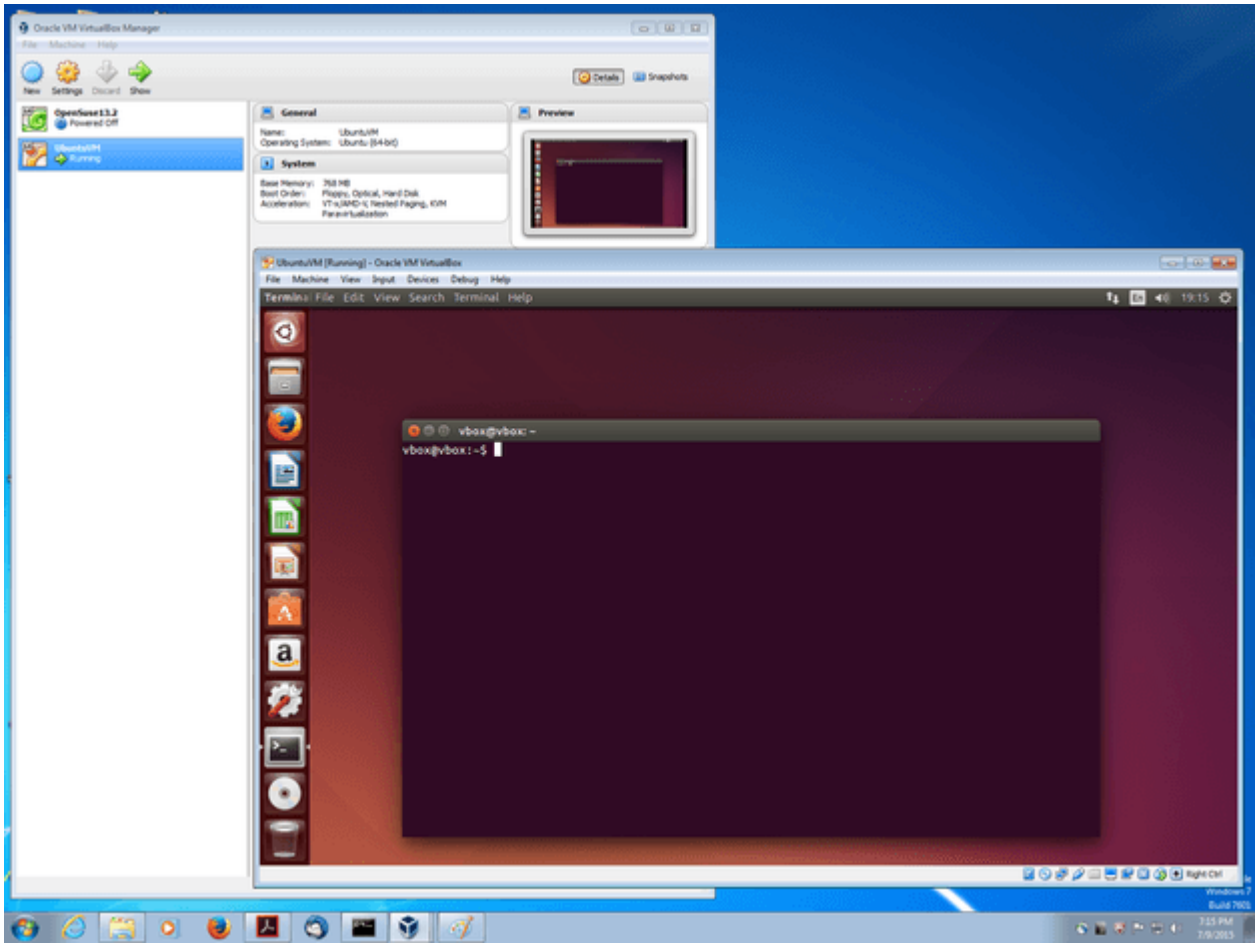
Getting started: how to run Bio-Linux as a VM

Here is a brief guide on how to set up a Virtual Machine on your PC to simulate a Linux environment with several bioinformatics tools.

Downloading VirtualBox

VirtualBox is a free and powerful cross-platform VM manager found at https://www.virtualbox.org/.

1. Ensure you have at least 40GB of free disk space.
2. Download and install the appropriate version of VirtualBox using the link above.
3. Follow the installation instructions.
4. Wait before starting any new VM.



VirtualBox 5.0 for Windows. Within VirtualBox Ubuntu 14.04 is running.

For further info on how to setup a VM on/with whichever OS you like, please refer to the manual (also enclosed to this email).

Downloadin Bio-Linux 8 as an OVA file

In order to minimize the number of tools we need to manually set up for our training, we choose to work with Bio-Linux 8, a free bioinformatics workstation platform that can be installed on anything from a laptop to a large server, or run as a virtual machine. Bio-Linux 8 adds more than 250 bioinformatics packages to an Ubuntu Linux 14.04 LTS base, providing around 50 graphical applications and several hundred command line tools. You can find more information on it here[10].

⚠ Bio-Linux is a 64-bit operating system. Virtually all modern PC processors support 64-bits, even if you have 32-bit Windows installed. As a rule of thumb, if you have more than 1 processor core you will have 64-bit support. See: https://www.virtualbox.org/manual/ch03.html#intro-64bitguests.

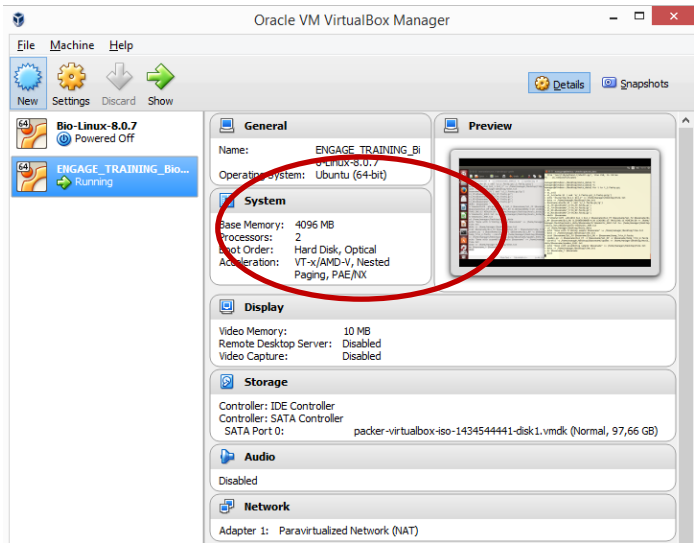
For our purposes, you should download the Bio-Linux 8 OVA file from http://nebc.nerc.ac.uk/downloads/bio-linux-8-latest.ova. The OVA file is designed for use with VirtualBox but should also work with similar systems like VMWare and Parallels.

Setting up your VM instance

To setup Bio-Linux 8 for VirtualBox:

1. Start VirtualBox
2. Select Import Appliance from the File menu and import the .ova file (don't worry that it says you need an OVF file) [NOTE: this step may take several minutes to perform…]
    a. When importing the appliance, select the option to reinitialize the MAC addresses of network cards.
3. Start the VM
4. If you see a log-in screen, log in as user **manager** with password **manager**.

Once this is working, you can delete the .ova file to save space. See the VirtualBox docs for more details including how to share folders (also detailed in the next paragraph) and hardware. You will also want to adjust hardware settings such as CPU, RAM and video acceleration settings to suit your hardware, by tuning the parameters of the "System" tab of your VM (when it is not running).



For example, on a Windows 8.1 machine with

- Intel i5-5200U CPU @ 2.20GHz 2.20GHz processor
- 8.00 GB RAM memory
- 64bit operating system, x64 processor

[10] Note, however, that this project is no more funded/developed and therefore there might be a better long-term choice to setup a Linux/Ubuntu based machine where you can install all the tools you need.

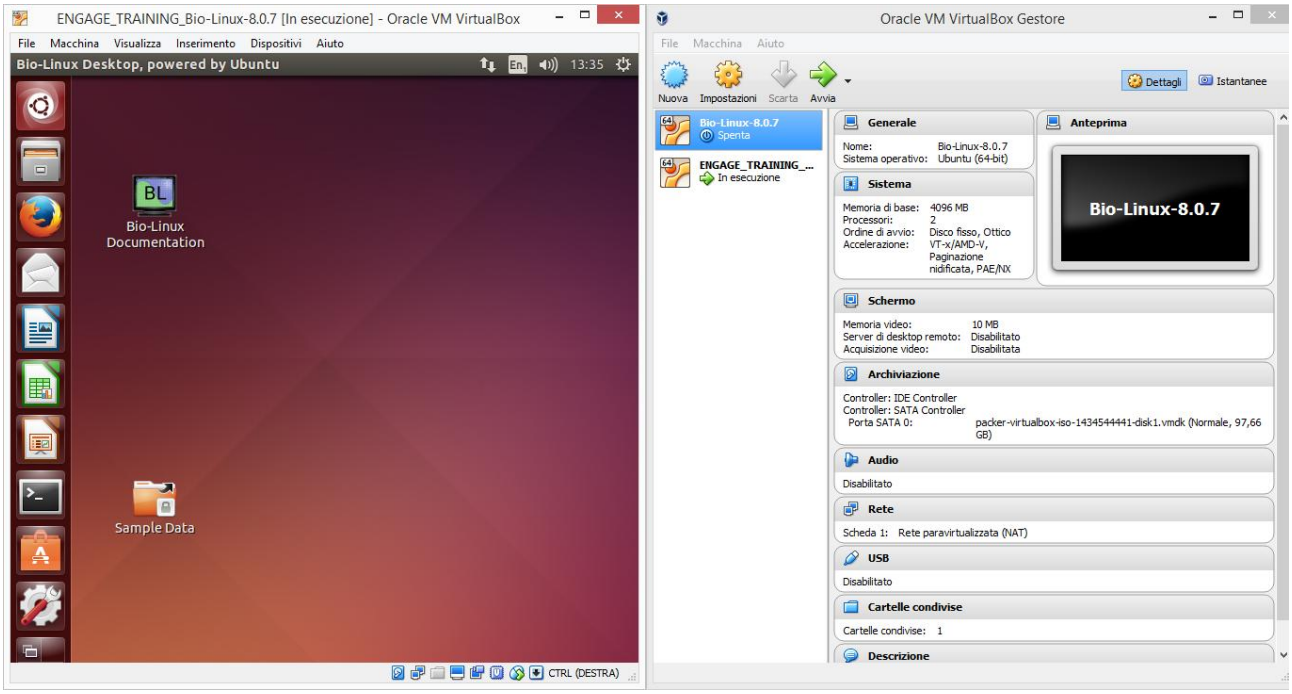
we suggest the following settings:

- Memory: 4096 MB
- 2 CPUs
- 10 MB video memory

or, more generally, we suggest to set both memory and CPUs values at half the value of your actual system and never below 2GB of memory.

Now you're ready to start your Biolinux VM!
For a list of all the tools included in this release of Bio-Linux, see this page.

NOTE: You should treat the VM as a real machine for security purposes and apply all system security updates in a timely manner. The default manager password is, clearly, not secure. This might not be a problem because by default nobody can access the Linux VM unless they have direct access to your computer, but if you open up the network settings (eg. by adding port forwarding rules) then you must secure the account with a strong password or else take other steps to limit remote access. Ideally enforce key-only access via SSH.



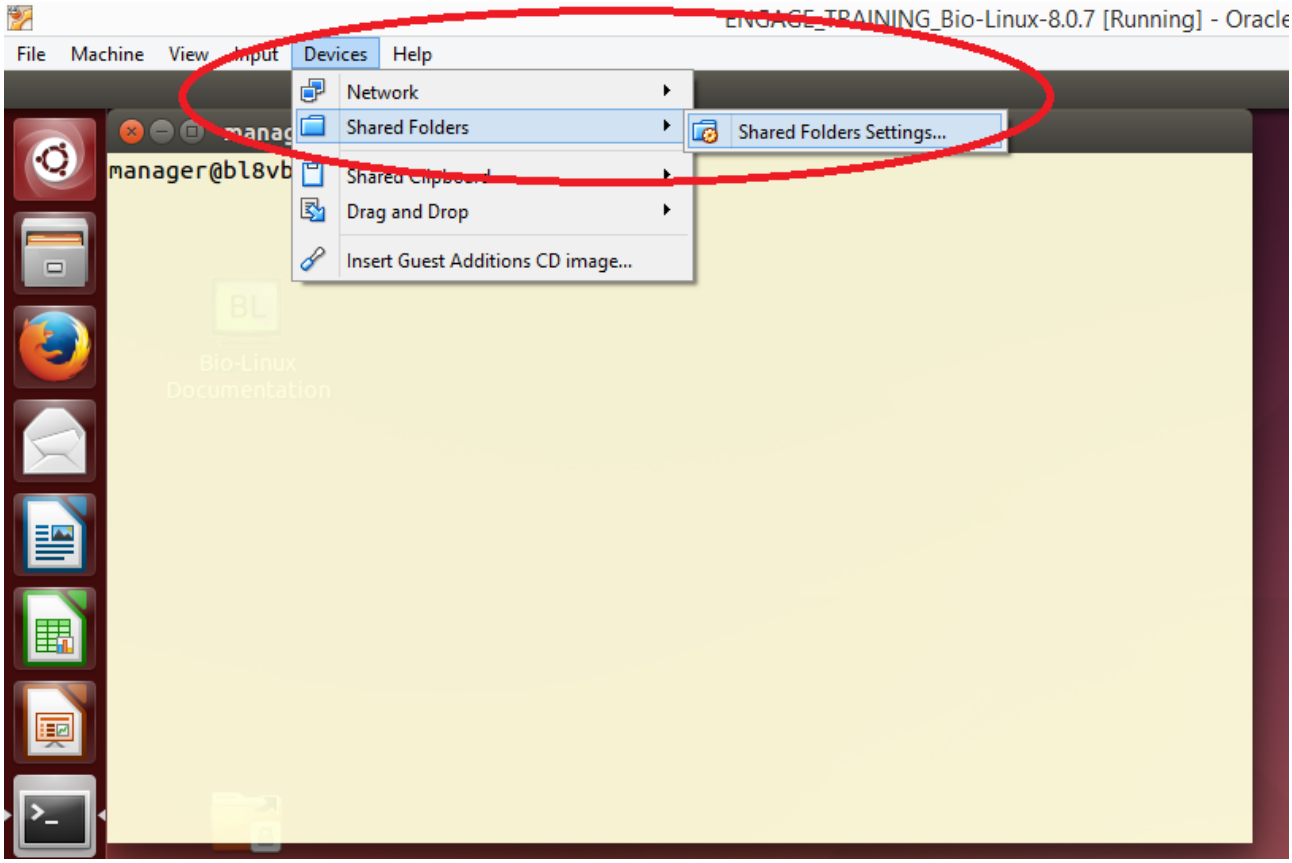
A screenshot of my Bio-Linux VM instance

Setting up a shared folder between your real machine and the VM

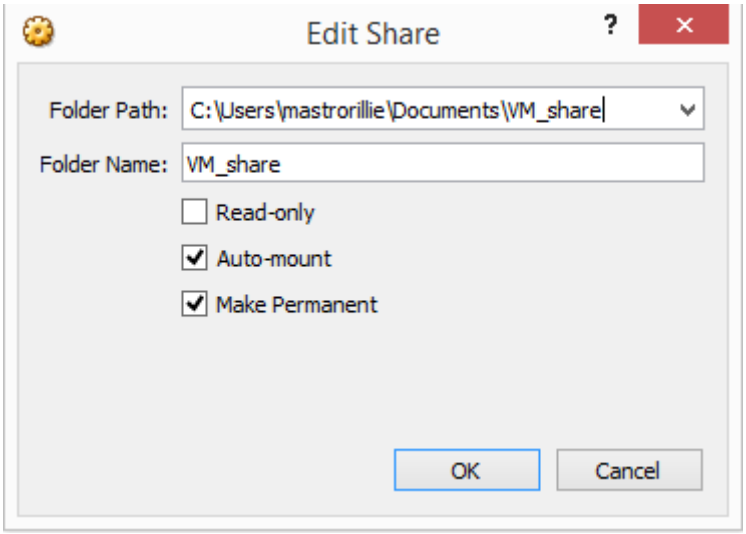
It is sometimes very useful to have the chance of sharing files between your real machine and the VM. With the "shared folders" feature of VirtualBox, you can access files of your host system from within the guest system. This is similar how you would use network shares in Windows networks – except that shared folders do not need require networking, only the Guest Additions. Shared Folders are supported with Windows (2000 or newer), Linux and Solaris guests. Shared folders must physically reside on the host and are then shared with the guest, which uses a special file system driver in the Guest Addition to talk to the host. For Windows guests, shared folders are implemented as a pseudo-network redirector; for Linux and Solaris guests, the Guest Additions provide a virtual file system. To

share a host folder with a virtual machine in VirtualBox, you must specify the path of that folder and choose for it a "share name" that the guest can use to access it. Hence, **first create the shared folder on the host** (e.g. we will refer to a folder called VM_share that I have in the Documents folder on my Win machine); then, within the guest, connect to it. In order to set an existing folder (on the host) as shared (with the VM)

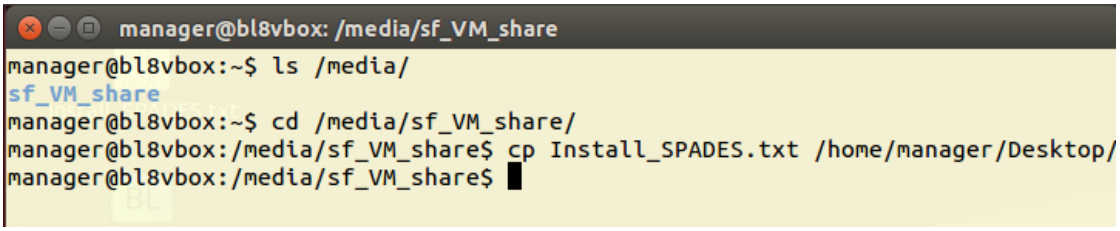
- Start your VM
- Go to Devices > Shared Folders > Shared Folders Settings…



- Use  to add a shared folder
- Navigate to the folder path
- Tick the options "Auto-mount" and "Make Permanent"
- Restart your virtual machine to see the changes.



- This will link your VM_share folder between the real and virtual machine, by putting it into the /media folder on Bio-Linux. Note that all shared folders will have "sf_" as a prefix.
- Now you can move to that directory (either from command line or from file explorer GUI) and copy files from it to have them locally on the VM.



Installing some tools from command line

Now all is set up to start working on your VM. If you want, you can try installing these two tools (which are not included in the Bio-Linux 8 release and that we will be using a lot during the training) directly from the command line. Please make sure you have internet connection available and open a terminal window to follow the instructions below

Trimmomatic

Trimmomatic is a flexible read trimming tool for Illumina NGS data. It is a Java-based tool, so first of all check if you have it installed on your VM by typing

```
which java
```

(default output should be /usr/bin/java). Then get trimmomatic by typing

```
sudo apt-get install trimmomatic
```

and **insert the password "manager"**. Once the installation is completed, you should be able to find it by typing

```
which TrimmomaticPE
```

To get usage information, just type

```
man TrimmomaticPE
```

on the command line.
To use Trimmomatic, you need to retrieve the ADAPTERS files (fasta format).
Run

```
#### GET THE ADAPATERS FOR TRIMMOMATIC
cd /usr/local/bioinf
sudo wget \ http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.36.zip
```

(note that the last three lines is actually one command only).
Then type

```
sudo unzip Trimmomatic-0.36.zip
```

to extract the files. A usage example for Trimmomatic would be

```
#### RUN TRIMMOMATIC ON ONE SAMPLE

/!\ you should de locate one folder above the sample!
training_set
|
| _ _ sample1
| _ _ sample2
| _ _ sample3
| _ _ sample4
| _ _ ...
| _ _ sample9

TrimmomaticPE -phred33 sample1/sample1.raw.R1.fastq.gz \ sample1/sample1.raw.R2.fastq.gz
sample1/sample1.raw.process.R1.fastq.gz\ sample1/sample1.raw.orphans.R1.fastq.gz
\ sample1/sample1.raw.process.R2.fastq.gz \ sample1/sample1.raw.orphans.R2.fastq.gz
\ ILLUMINACLIP:/usr/local/bioinf/Trimmomatic-0.36/adapters/NexteraPE-PE.fa:2:30:10:8:true
LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 \ MINLEN:50
```

NOTE: Remember that your output files should always be in the format:

sample1_R1_processed	sample1_R1_orphans
sample2_R2_processed	sample2_R2_orphans

You can retrieve more information (all the explanation for options meaning and why/how to set them) from Anais's presentation.

SPAdes

SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines.
To get it, open the terminal and type

```
wget http://cab.spbu.ru/files/release3.11.0/SPAdes-3.11.0-Linux.tar.gz
```

Move it to the bin folder

```
sudo cp SPAdes-3.11.0-Linux.tar.gz /usr/local/bin
```

if password is required, type "manager". Move to the selected folder and uncompress the file

```
cd /usr/local/bin
sudo tar –xzf SPAdes-3.11.0-Linux.tar.gz
```

[Optional] Create a soft link to the folder, so you don't have to change much if you install a newer version later on:

```
sudo ln –s SPAdes-3.11.0-Linux/ spades
```

Add the folder to the path by modifying the .zshrc file (if your command line interpreter is zsh) or your /etc/profile file (if your command line interpreter is bash)[11]

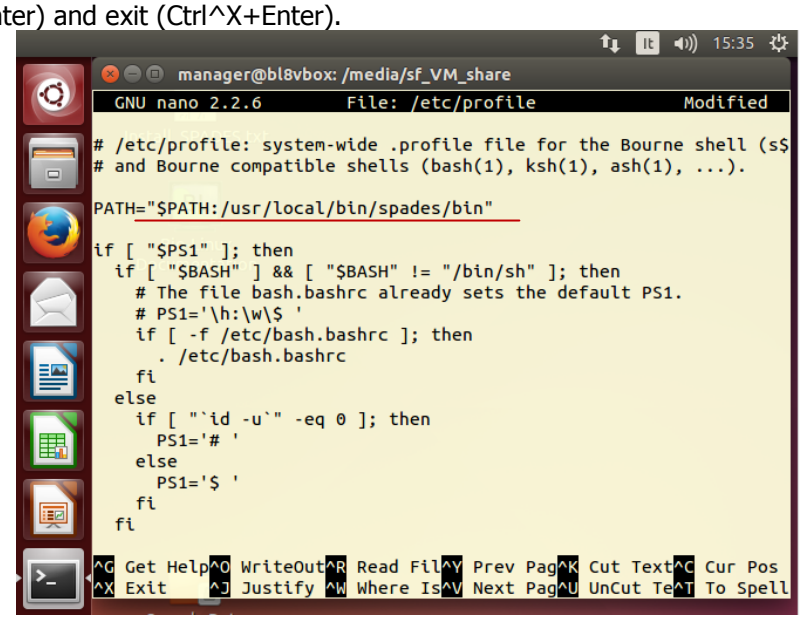
[11] In order to test which interpreter you are using, write echo $0 on the command line. If your result is zsh and you wish to change it to bash, just type "chsh -s /bin/bash" on the command line and restart the VM.

```
sudo nano ~/.zshrc
```

add the line in the header of the file (see next screenshot).

```
PATH="$PATH:/usr/local/bin/spades/bin"
```

Save (Ctrl^O+Enter) and exit (Ctrl^X+Enter).



Screenshot of the .zshrc file. Please insert the PATH command right after the comments (lines starting with #) and ignore the rest of the file content.

In order to see the changes to the path without restarting the VM, re-type in the command line

```
export PATH="$PATH:/usr/local/bin/spades/bin"
```

For testing purposes, SPAdes comes with a toy data set (reads that align to first 1000 bp of E. coli). To try SPAdes on this data set, run from command line:

```
spades.py --test
```

If the installation is successful, you will find the following information at the end of the log:

```
===== Assembling finished. Used k-mer sizes: 21, 33, 55
 * Corrected reads are in spades_test/corrected/
 * Assembled contigs are in spades_test/contigs.fasta
 * Assembled scaffolds are in spades_test/scaffolds.fasta
 * Assembly graph is in spades_test/assembly_graph.fastg
 * Assembly graph in GFA format is in spades_test/assembly_graph.gfa
 * Paths in the assembly graph corresponding to the contigs are in spades_test/contigs.paths
 * Paths in the assembly graph corresponding to the scaffolds are in spades_test/scaffolds.paths
======= SPAdes pipeline finished.
========= TEST PASSED CORRECTLY.
SPAdes log can be found here: spades_test/spades.log
Thank you for using SPAdes!
```

Quast

Quast is a quality assessment tool for measuring the quality of your genome assembly. It is particularly useful because it can generate a table comparing different metrics of your genome assemblies.

To download the tool, run:

```
wget https://downloads.sourceforge.net/project/quast/quast-4.5.tar.gz
sudo cp quast-4.5.tar.gz /usr/local/bin
cd /usr/local/bin
sudo tar -xzf quast-4.5.tar.gz
```

```
echo "PATH=\"$PATH:/usr/local/bin/quast-4.5\"" >> ~/.zshrc
export PATH=\"$PATH:/usr/local/bin/quast-4.5\"
```

Let's analyze what these lines are doing:

1. get the compressed installation file from internet
2. copy the compressed file into the /usr/local/bin folder; you have to use sudo to have the administrator permissions to copy into this folder
3. change directory to /usr/local/bin
4. uncompress your compressed file
5. update your config file so to add the quast folder to your PATH variable
6. update your PATH on the fly to avoid rebooting your machine.

Now that you have quast at hand, you can use it with a list of contig files to compare their qualities.

References

https://www.virtualbox.org/

http://environmentalomics.org/whats-new-in-bio-linux-8/

http://www.usadellab.org/cms/index.php?page=trimmomatic

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/

http://cab.spbu.ru/software/spades/

http://quast.sourceforge.net/

https://www.ncbi.nlm.nih.gov/pubmed/22506599

Appendix E – Benchmarking of *de novo* assembly tools: SPAdes 3.9 vs Velvet 1.2

Report number	#1
Responsible	Pimlapas Leekitcharoenphon (DTU) and Maria Borowiak (BfR)
Other partners/institutions involved	-
Benchmarking launched (date)	May 2016
Deliverable due (date)	Due: May 2016 Delivered: August 2016

Purpose of the benchmarking exercise

The purpose of this benchmarking exercise was to evaluate and compare the performance of the mostly used *de novo* assembly tool, i.e. Velvet, and the newer introduced *de novo* assembly tool, SPAdes.

Tools included in the benchmarking exercise

De novo assembly tools; Velvet 1.2 with default parameters (Assembler-1.2 implemented in the tool Bacterial Analysis Pipeline - Batch Upload (https://cge.cbs.dtu.dk/services/cge/)) and SPAdes 3.9 (http://cab.spbu.ru/software/spades/) with default parameters in careful mode. Both tools were run using different k-mer sizes and the assembled genome was set to pick up from the best k-mer size.

Species and/or genomes included

50 *Salmonella enterica* subsp. *enterica* serovar Paratyphi B *d*Ta+ (S. Java) isolates were tested. DNA from bacterial cells was isolated from liquid cultures using the PureLink® Genomic DNA Mini Kit (Invitrogen, Carlsbad, CA, USA). Sequencing libraries were prepared with the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. Paired-end sequencing was performed in 2 × 300 cycles on the Illumina MiSeq benchtop using the MiSeq Reagent v3 600-cycle Kit (Illumina). Further details related to the included genomes can be found at the end of this report in Table E.2 and in Supplementary Table 2 (Annex B).

Results

Overall assembly quality

Sequencing raw data without trimming was assembled using either Velvet or SPAdes assembly tools. Analysis of contigs using ContigAnalyzer-1.0, implemented in the Bacterial Analysis Pipeline - Batch Upload (https://cge.cbs.dtu.dk/services/cge/), revealed that the mean number of contigs is lower and the mean N50 value (median contig size of a genomic assembly) is higher in the genomes assembled using SPAdes (see Table E.1 and Figure E.1). The observed mean genome size however is similar for both assembly types.

Table E.1: Assembly quality analysed using ContigAnalyzer-1.0

		Spades	Velvet
Contig number	**mean**	100	249
	min	51	144
	max	181	376
	sd	30	55
N50	**mean**	176,144	57,148
	min	53,662	26,926
	max	393,606	146,576
	sd	93,110	23,786
Assembled genome size	**mean**	4,924,464	4,872,591
	min	4,663,179	4,505,678
	max	5,076,872	5,027,353
	sd	101,043	121,670

To further assess the quality of the assemblies, the Multi Locus Sequence Type (MLST) and antibiotic resistance genes were analysed.

Results regarding MLST identification

Analysis of the obtained assemblies regarding the Multi Locus Sequence Type (MLST) was performed using the tool MLST 1.6 (https://cge.cbs.dtu.dk/services/cge/). MLST types (based on the Enterobase scheme, https://enterobase.warwick.ac.uk) could be predicted in 100% of the SPAdes assembled and in 94% of the Velvet assembled genomes.

Results regarding the identification of resistance genes

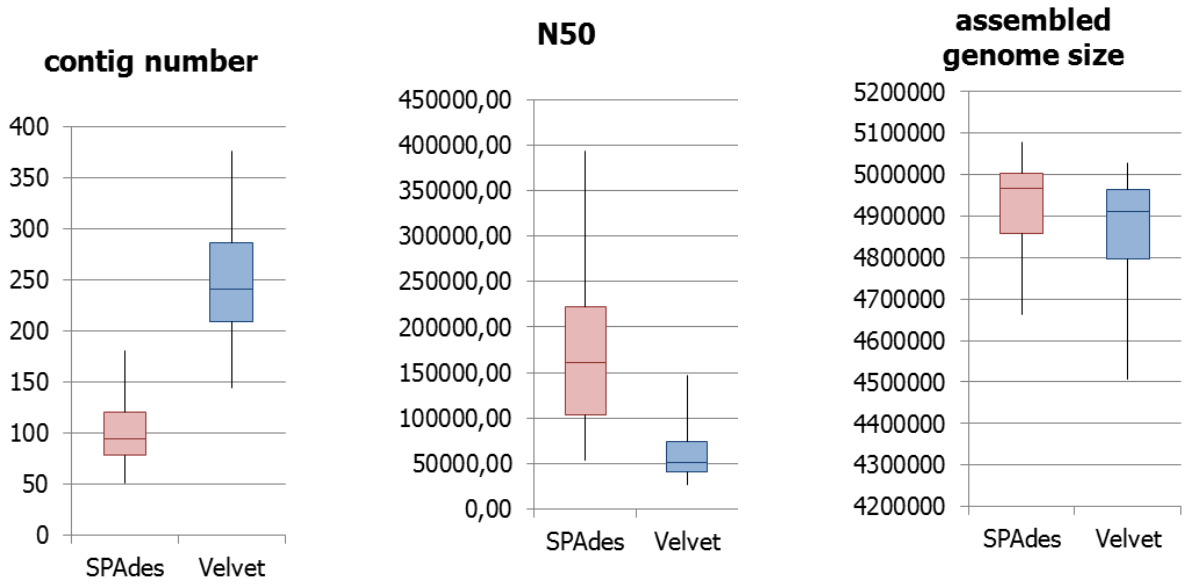
Antimicrobial resistance patterns derived from MIC values (obtained by broth microdilution method following CLSI guidelines, and using the EUCAST epidemiological ECOFFs; testing conditions applied to the individual samples depend on the year the isolate was collected and are listed in Supplementary Table 2 (Annex B)) were compared with the ResFinder2.1 (https://cge.cbs.dtu.dk/services/cge/) output (AMR genes detected) for *de novo* assembled sequence data (see Supplementary Table 2 (Annex B)).

Concordance between genotypic and phenotypic resistance data (for detailed results see also Supplementary Table 2 (Annex B)):

- In 35/50 cases the phenotypic resistance profile could be explained with genes found using Velvet as assembler.
- In 38/50 cases the phenotypic resistance profile could be explained with genes found using SPAdes as assembler.
- In 12/50 cases the phenotypic resistance profile could not be explained with genes found using either SPAdes or Velvet for assembly, one or more genotypic resistance determinants were missing.
- In 5/50 cases resistance genes conferring resistance to aminoglycosides which were not expected based on phenotypic resistance data were found in both genome assemblies.
- In 7/50 cases additional resistance genes which were not expected based on the phenotypic resistance profiles were found in the genomes assembled using SPAdes. This involves *aac(3)*-VIa-like genes (6 cases) and *erm(B)* (1 case).

Conclusions

All in all, SPAdes assembled genomes showed longer contigs and therefore higher N50 values. This seems to lead to an improved detection of MLST genes. Moreover, "missing" resistance genes, i.e. those absent from genomes assembled using Velvet, could be identified when using SPAdes for genome assembly. Nevertheless, there is a huge number of cases where not all expected genetic resistance determinants were identified. This can be caused by loss of resistance plasmid during storage and culturing or emergence of unknown resistance mechanisms and chromosomal point mutations which could not be identified using the ResFinder2.1 tool. Additional identification of streptomycin resistance determinants, which were not expected based on phenotypic data, are likely to be caused by incorrectly determined MIC values or changes regarding break points and test panels. For better comparison of the data, isolates with contradicting phenotypical and genotypical results should be subjected to MIC retesting. In case of the *aac(3)*-VIa-like genes and the *erm(B)* that were detected in 7 SPAdes assembled genomes, further analysis of the respective contigs revealed that all of them showed a low coverage. These contigs might have been derived from the assembly of low level read contaminations from other samples which might have led to the false positive detection of genotypic resistance determinants. Including low coverage contigs caused by read contamination in the assembled genomes might be a disadvantage of SPAdes. Additional filters should be applied to remove low coverage contigs.



Graphical representation of overall assembly quality parameters including contig numbers, N50 values and genome sizes of genomes assembled with either SPAdes or Velvet.

Figure E.1: Overall assembly quality

Table E.2: List of Strains (see also Supplementary Table 2 in Annex B)

	Spades			Velvet		
sample_name	genome_size	contigs	n50	genome_size	contigs	n50
03-02917	4674923	150	70852	4505678	360	26926
06-02242	4762839	157	88213	4633625	335	30290
07-01597	4663179	64	225719	4577464	213	51461
08-00436	4896492	118	119248	4797832	314	35933
08-00436	4967144	79	247068	4940435	222	85291
08-00844	4970846	91	213767	4941452	230	58722
08-00955	4965087	100	155361	4876611	278	43853
08-03422	4955841	120	137558	4876128	293	44300
09-02362	4871450	88	174043	4804866	227	53647
09-02946	5034312	91	225719	5027353	200	85169
09-02986	4954613	146	103875	4844786	337	30225
09-03610	4926660	97	164864	4918582	205	74734
09-04431	4962053	88	187927	4919965	239	51201
10-03145	4915801	181	53662	4754476	354	31818
10-03460	4818113	63	368622	4788341	346	34646
10-04072	4913494	122	82860	4833220	270	46531
10-04072	4909537	81	165445	4883184	192	76987
10-05043	4991716	172	68232	4888669	376	30963
11-01176	4782703	113	124638	4720850	271	44563
11-01525	4972448	92	184458	4962843	183	103705
11-02165	4966007	86	166565	4907581	242	44379
11-03654	5012273	83	173228	4986940	224	54113
11-03655	5011129	72	393606	4969447	222	56233
11-03656	5013701	86	206171	4995442	189	96509
11-04054	4897942	140	77220	4859167	290	40921
11-04056	4912808	90	165788	4873888	238	62293
11-04559	5014967	69	368674	5007664	144	146576
12-00555	5007211	115	94646	4855916	287	41914
12-01208	5016473	93	157181	4958028	248	48398
12-02541	4707937	128	93229	4634546	285	37900
12-02857	4774719	124	96314	4678889	302	35324
13-SA02194	4970145	75	385587	4943543	167	90284
13-SA02281	5008432	120	96736	4968954	303	38101
13-SA02283	4983075	68	253523	4968764	199	74230
13-SA02300	4986840	98	147698	4964091	248	46308
13-SA02435	4982663	104	121634	4949929	236	53586
13-SA02656	5076872	124	100656	5021983	285	45019
13-SA02788	4967735	80	192581	4948003	216	56954
14-SA00333	5010528	62	231654	4995680	210	83814
14-SA00775	4813906	109	103703	4772262	248	38773
14-SA00777	4987252	95	134015	4950980	259	51549
14-SA00918	4964052	96	121174	4914842	252	44954
14-SA01149	5013356	60	368866	4999641	185	93083
14-SA02536	5014878	69	275055	4998872	200	79253
14-SA02741	5009213	122	128575	4941385	287	47436
14-SA02860	4993807	116	131105	4961677	234	52422
15-SA00146	4776301	136	62450	4722696	267	35590
15-SA01434	4807642	67	172824	4795487	174	79619
15-SA01523	4805362	51	392833	4797115	175	82555
15-SA02829	4806710	64	231776	4789754	213	58324

Appendix F – Benchmarking of genotypic *Salmonella* serotype prediction (general)

Report number	#2
Responsible	Anthony Underwood (PHE)
Other partners/institutions involved	Lauren Cowley (PHE), Rolf Sommer Kaas (DTU), Pimlapas Leekitcharoenphon (DTU), Rob Davies (APHA), Mirko Rossi (University of Helsinki/ INNUENDO), Kathie Grant (PHE), Liljana Petrovska (APHA), Rene S. Hendriksen (DTU), Susanne Karlsmose Pedersen (DTU)
Launch date	Nov 2016
Deliverable date	Dec 2016

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools for predicting the *Salmonella* serotype. Some EC regulations require the use of conventional serotyping methods. This could influence the need and velocity in the implementation of NGS for animal and food surveillance.

Tools benchmarked

Benchmarking by determining serotype genotypical using the following tools with default parameters:

1) MOST (PHE tool) run by DTU

2) SalmonellaTypeFinder 1.4 run by PHE

3) SeqSERO 1.2 stand-alone tool run by APHA and by PHE (as part of SalmonellaTypeFinder)

4) SISTR v1.0.1 run by INNUENDO (https://lfz.corefacility.ca/sistr-app/)

Species/genomes included

Three datasets have been collected for this study (See below Table F.4 of "Tested serotypes", Annex A, Annex C). Strain selection was based on inclusion of a wide variation of serovars including commonly isolated serovars and rare serovars, seldom found (Table F.1).

The Animal and Plant Health Agency (APHA) collected 78 serotyped *Salmonella* isolates. The dataset included 78 serotypes of which all were rare serovars. Bacterial DNA was extracted using the MagNA Pure LC DNA Isolation Kit III (Roche) according to manufacturer's instructions and sequencing libraries were prepared using the NexteraXT sample preparation method for sequenced on the Illumina HiSeq platform with paired-end 2x125bp reads (http://www.illumina.com).

The National Food Institute at DTU collected 208 serotyped *Salmonella* isolates (these dataset were not sequenced under ENGAGE project). The dataset included 208 isolates from 87 serotypes, received from the project '100K Salmonella project' (external subproject lead by this affiliated partner, data not included in this report). Genomic DNA was using an Invitrogen Easy-DNA™ Kit (Invitrogen, Carlsbad,

CA, USA) and DNA concentrations were determined using the Qubit dsDNA BR assay kit (Invitrogen). The genomic DNA was prepared for Illumina pair-end sequencing using the Illumina (Illumina, Inc., San Diego, CA) NexteraXT® Guide 150319425031942 following the protocol revision C (http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html). A sample of the pooled NexteraXT Libraries was loaded onto a Illumina HiSeq reagent cartridge using HiSeq Reagent Kit v2. The libraries were sequenced using an Illumina Hiseq platform.

Public Health England (PHE) collected 500 serotyped *Salmonella* isolates. The dataset was selected to represent the serotypes that PHE receives routinely as a public health agency in the UK. It included 500 isolates from the PHE collection and representing 104 serotypes included in the PHE collection. DNA extraction of *Salmonella* isolates begins with a manual lysis using ATL buffer, Proteinase K and RNAase A (Qiagen, Hilden, Germany) (220µl, 20µl and 4µl respectively) before loading onto a Qiagen Qiasymphony SP for purification.

DNA quantification was performed using the Promega GloMax with the Invitrogen Quant-iT dsDNA Assay Kit (Broad range) (ThermoFisher Scientific, Waltham, Massachusetts, United States) according to the manufacturer's instructions. Genomic DNA was then processed using the NexteraXT® sample preparation method and sequenced with a standard 2x101 base protocol on a HiSeq 2500 Instrument in fast mode (Illumina, San Diego, CA, USA).

All selected isolates were serotyped phenotypically according to the WKLM scheme.

Table F.1: Strain providers, number of isolates and serotypes

DATASET	ISOLATES	SEROTYPES
APHA SERIES	78	78
DTU SERIES	208	87
PHE SERIES	500	104
TOTAL	**786**	**196***

*The number of serotypes is unique serotypes across the total dataset and therefore not a sum of the serotypes within each dataset.

All datasets were sequenced on an Illumina HiSeq.

Method

Four tools have been benchmarked in this study: Metric-Oriented Sequence Typer (MOST), SeqSero (Zhang et al., 2015), SalmonellaTypeFinder, The *Salmonella In Silico* Typing Resource (SISTR).

Availability of tools:

MOST: https://github.com/phe-bioinformatics/MOST

SeqSero: https://github.com/denglab/SeqSero

SeqSero web tool: http://www.denglab.info/SeqSero

SalmonellaTypeFinder: https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder/

SISTR: https://lfz.corefacility.ca/sistr-app/.

Briefly, about the three tools:

MOST is based on the first version of the tool "Short Read Sequence Typing" (SRST) (Inouye et al., 2012). MOST maps read data to MLST genes and infers an MLST type. The MLST type is subsequently looked up in a local MOST database that contains information on which serotypes that have been registered for the MLST type in question. The local database has been divided into two parts, one

database containing only information from PHE and another with information collected from EnteroBase (http://enterobase.warwick.ac.uk/)

SeqSero is doing *in silico* molecular serotyping. In the sense that it maps read data to a local database of the genes that causes the phenotype of the specific serotypes. SeqSero thereby infers the phases and translates the phase profile into a serotype.

SalmonellaTypeFinder is an attempt to merge the methods from the above tools. SalmonellaTypeFinder runs SeqSero and infers an MLST type using SRST2 (Inouye et al., 2014). The MLST type is subsequently looked up in a local database created from information in EnteroBase (this includes information from PHE) to determine which serotypes that have been registered for the particular MLST type. A serotype is then inferred from the MLST type based on the criteria that at least 3 registered isolates of the same serotype has been found with the MLST type in question, and at least 75% of the serotypes registered to the MLST type is identical. The final serotype is then found by comparing the serotype inferred by SeqSero and the serotype inferred by MLST. The serotype from SeqSero always takes precedence over the serotype inferred by MLST. If SeqSero reports several serotypes, the serotype (if any) agreeing with the MLST serotype is chosen.

SISTR is a bioinformatics platform for rapidly performing simultaneous *in silico* analyses for several leading subtyping methods on draft *Salmonella* genome assemblies. The serovar prediction module in the SISTR server utilizes O (somatic) and H (flagellar) antigen and/or serogroup-specific probes previously designed for our *Salmonella* Genoserotyping Array (SGSA), which provides serovar identification for 90% (n = 2,190) of serovars.

All isolates were trimmed using bbduk2 (part of the suit bbtools version 36.49) and *de novo* assembled using SPAdes. All isolates were analysed using all four tools. PHE ran the tools SalmonellaTypeFinder and SeqSero. SeqSero was run as a part of SalmonellaTypeFinder. DTU ran the tool MOST. The output from most is an array of all the serotypes registered to a specific MLST type, along with information on which were registered by PHE and which were not in EnteroBase. The authors of MOST (PHE) decided to predict a serotype by selecting the most commonly registered, by PHE, serotype for an MLST type. INNUENDO coordinator (University of Helsinki) ran SISTR.

Overall results

The results were divided into the serotype predictions that correlate with the expected serotype (Table F.2, Figure F.1) and those that do not correlate. The results that does not correlate has been further divided in to the predictions that give a different serotype than the expected (miscorrelation, Figure F.2), the predictions that yields no result (no prediction, Figure F.3), and the predictions that yield several possible serotypes, were the expected serotype is found among those (ambiguous, Figure F.4). For more detail of all the results, see Supplementary Table 3 in Annex C.

Table F.2: Serotype prediction results

	MOST	***SeqSero***	***SalmonellaTypeFinder***	***SISTR***
Correlation	668 (85%)	508 (65%)	669 (85%)	694 (88%)
No Correlation	118 (15%)	278 (35%)	117 (15%)	92 (12%)
- *Miscorrelation*	33 (4%)	22 (3%)	26 (3%)	65 (8%)
- *No prediction*	85 (11%)	34 (4%)	34 (4%)	8 (1%)
- *Ambiguous*	0 (0%)	222 (28%)	57 (7%)	19 (2%)

7 miscorrelations (0.9% of all isolates) were identical across all four tools, meaning that all the tools agreed upon the predicted serotype. The relatively low correlation for APHA dataset was due to the fact that APHA dataset consist of rare serotypes (Table F.3).

Presented below is the correlation to each of the three datasets.

Table F.3. Correlation result for different series of data:

	MOST	SeqSero	SalmonellaTypeFinder	SISTR
APHA	17 (22%)	44 (56%)	46 (59%)	35 (45%)
DTU	169 (81%)	155 (75%)	191 (92%)	192 (92%)
PHE	482 (97%)	309 (62%)	432 (86%)	467 (93%)

Conclusion

The results of this benchmarking study clearly demonstrate that serotyping using NGS data is a very feasible option. The tool with highest correlation, SISTR, gets 88% correlation with the conventional serotyping (Figure F.5), and this is a conservative number, considered none of the isolates have be retested, to ensure correct serotyping.

The miscorrelation rate, cases were the tools predicted a different serotype than the expected, were 3-8% in this study. Additionally, at least half of these miscorrelations are heavily suspected to be mistakes in the conventional serotyping. Interestingly, the tool with highest correlation also seems to have the highest miscorrelation. It is not possible from this study to conclude why these miscorrelations happened, but the tools are under constant development and the errors made by the tools are decreasing with each new release.

Such a low miscorrelation rate would probably be hard to achieve for most labs that does conventional serotyping.

Three of the four tools archive similar scores with a correlation rate between 85-88% and "no correlation" rates between 12-15%. It is important to note that the lowest scoring tool "SeqSero" is an essential part of the higher scoring tool "SalmonellaTypeFinder".

Additional notes

It is recommended to serotype the isolates were the predictions from the tools disagree with the expected serotype. This is especially important for the isolates where all tools have identical miscorrelations. Additionally, the different sequence quality and sequencing processing may have an effect on the results.

References

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI and Deng X, 2015. Salmonella serotype determination utilizing high-throughput genome sequencing data. Journal of Clinical Microbiology, 53(5):1685-1692. doi: 10.1128/JCM.00323-15

Inouye M, Conway TC, Zobel J and Holt KE, 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics, 13, 338. doi: 10.1186/1471-2164-13-338

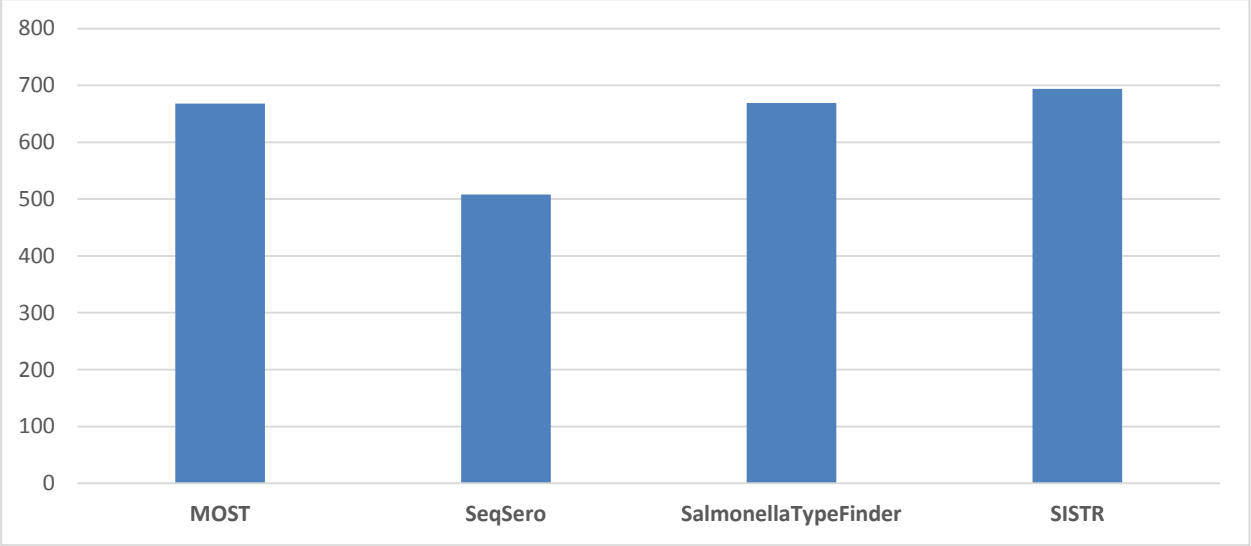
Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J and Holt KE, 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine, 6(11):90. doi: 10.1186/s13073-014-0090-6

Table F.4: Tested serotypes

Serotype	count	Serotype	count	Serotype	count
35:z10:-	1	Butantan	1	Hadar	7
38:k:-	1	Canastel	1	Haifa	8
4,12:d:-	1	Cerro	8	Havana	10
9,46:z45:-	1	Chandans	1	Heidelberg	7
Aberdeen	4	Chester	6	Hiduddify	1
Abony	5	Chicago	1	Hvittingfoss	5
Adelaide	5	Claibornei	1	Ibadan	5
Agama	5	Coeln	6	II 1,4,12:z29:e,n,x	1
Agbeni	5	Coleypark	1	II 16:m,t:[z42]	1
Ago	5	Colindale	7	II 21:z10:z6	1
Agona	6	Concord	8	II 55:k,z39:1	1
Agoueve	1	Corvallis	9	II 58:d:z6	1
Ajiobo	5	Derby	10	IIIa 41:z4,z23:-	1
Alachua	5	Dublin	8	IIIa 47:z4,z23:-	1
Albany	7	Dugbe	1	IIIa 51:g,z51:-	1
Altona	4	Durham	6	IIIa 56:z4,z23:z32	1
Amager	5	Ealing	3	IIIb 61:1,5,7:-	1
Amsterdam	2	Eastbourne	5	IIIb 65:z10:e,n,x,z15	1
Anatum	8	Elisabethville	1	Indiana	6
Anfo	1	Emek	6	Infantis	10
Ank	2	Enteritidis	24	Isangi	2
Apapa	2	Falkensee	1	Istanbul	1
Augustenborg	1	Fischerkietz	1	Itami	1
Bardo	1	Florida	1	Ituri	1
Bareilly	8	Fluntern	4	IV 48:g,z51:-	1
Bergen	1	Freetown	1	IV 50:g,z51:-	5
Bispebjerg	4	Fresno	1	Jangwani	5
Blockley	6	Friedenau	1	Javiana	7
Bonariensis	1	Gaminara	5	Jukestown	1
Bonn	3	Georgia	1	Kambole	1
Bovismorbificans	7	Give	8	Karachi	1
Braenderup	7	Glostrup	2	Kedougou	12
Brandenburg	7	Godesberg	1	Kentucky	12
Bredeney	7	Goldcoast	5	Kenya	5
Khami	3	Mpouto	1	Singapore	2
Kibi	1	Muenchen	8	Solt	1
Kimuenza	1	Muenster	6	Stanley	10
Kingston	5	Nagoya	1	Stanleyville	5
Kisangani	3	Napoli	5	Stockholm	1
Kisarawe	1	Newport	11	Takoradi	6
Kokomlemle	1	Nima	5	Tees	1
Kottbus	6	Nottingham	3	Telelkebir	6
Kuessel	1	Offa	1	Teltow	1
Landwasser	1	Ohio	5	Tennessee	5
Lexington	1	Omifisan	1	Thompson	7
Lille	1	Onireke	0	Toricada	1
Litchfield	7	Oranienburg	7	Typhi	5
Liverpool	1	Oritamerin	1	Typhimurium	28
Livingstone	6	Oslo	5	Uganda	3
London	7	Panama	8	Umbilo	5
Madjorio	1	Paratyphi A	5	Vejle	1
Malstatt	1	Paratyphi B	4	Vinohrady	1
Manchester	3	Paratyphi B var Java	7	Virchow	12
Manhattan	2	Pomona	1	Virginia	1

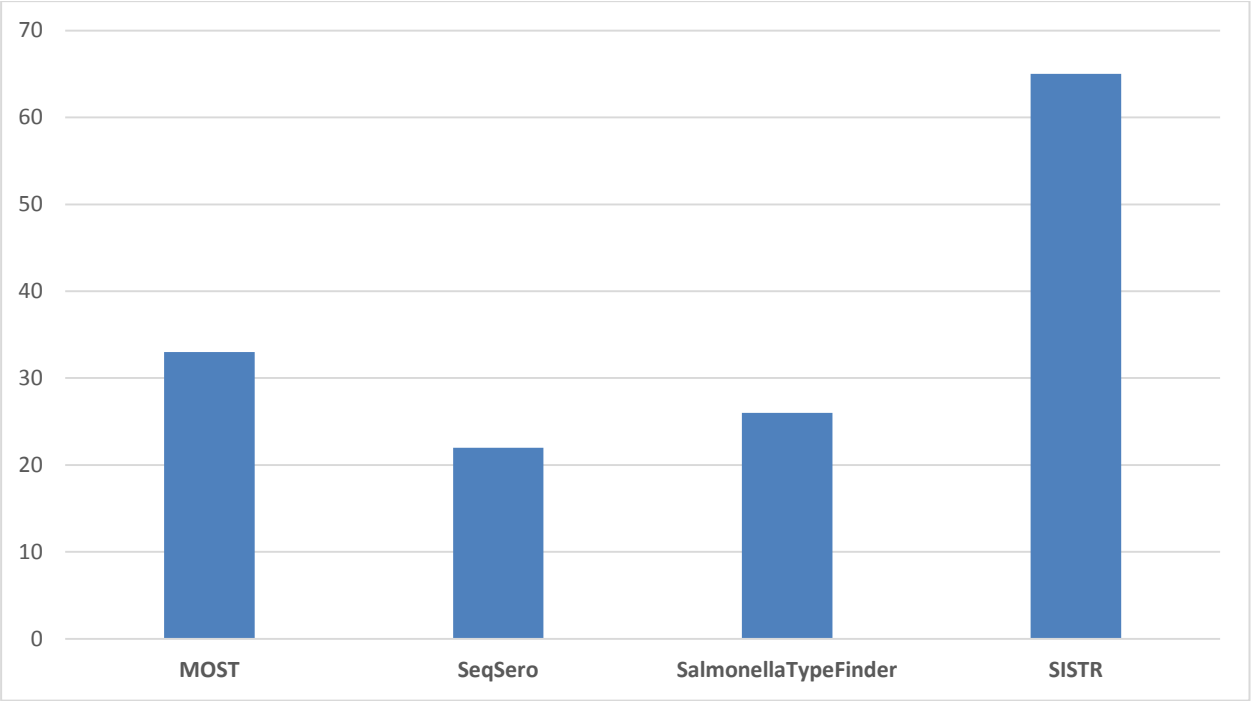
Serotype	count	Serotype	count	Serotype	count
Matopeni	1	Poona	7	Vitkin	4
Mbandaka	11	Potsdam	8	Vogan	1
Meleagridis	5	Putten	1	Wangata	1
Mgulani	1	Reading	2	Waycross	1
Mikawasima	5	Richmond	5	Weltevreden	8
Minnesota	6	Rissen	9	Westhampton	1
Mishmarhaemek	1	Rubislaw	5	Widemarsh	1
Mississippi	5	Saintpaul	8	Wilhelmsburg	1
Moero	1	Sandiego	6	Wippra	1
Monschaui	7	Schwarzengrund	8	Worthington	2
Montevideo	8	Senftenberg	8		
Morningside	1	Shipley	1		

Note: Information on the isolates included in this benchmarking analysis is available in Annex C.



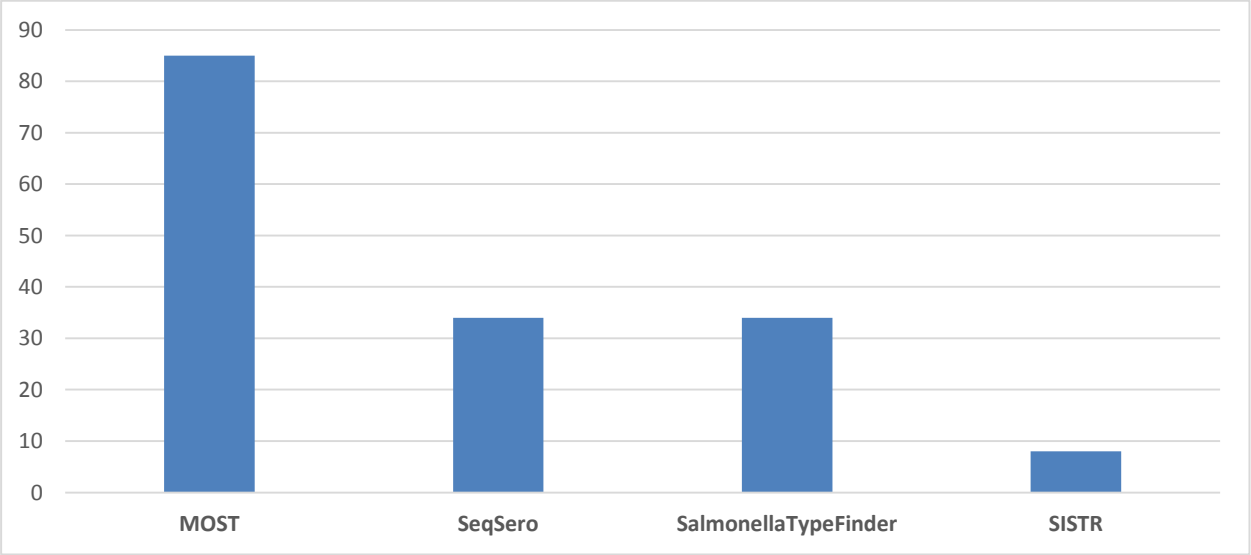
Y-axis represents number of isolates that serotype predictions correlate with the expected serotype.

Figure F.1: Correlations



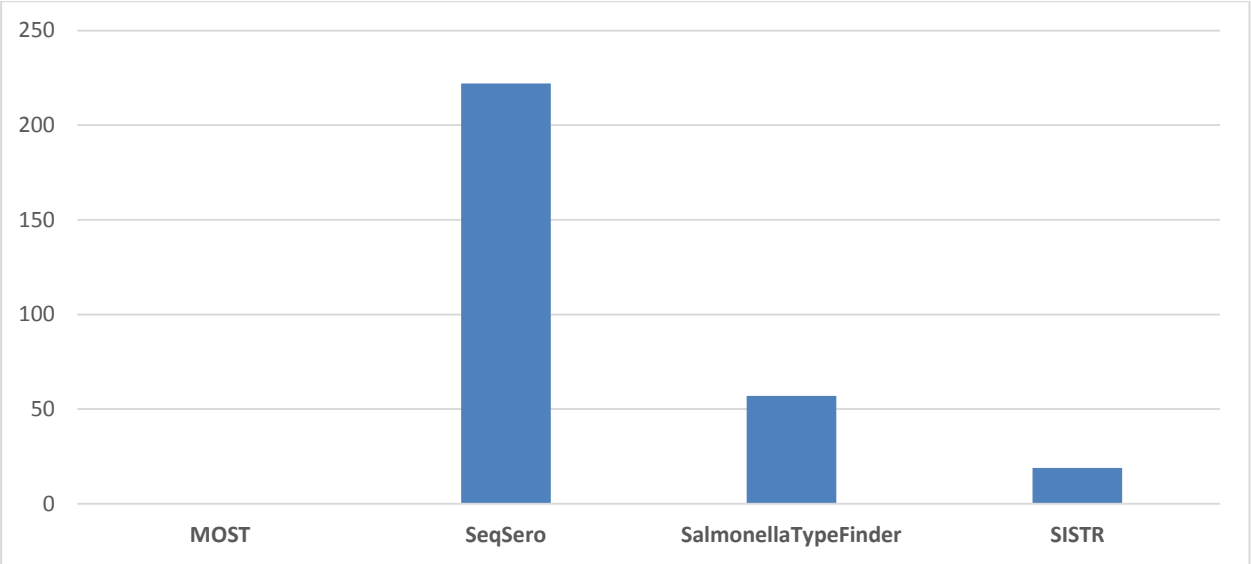
Y-axis represents number of isolates that serotype predictions give a different serotype than the expected.

Figure F.2: Miscorrelation



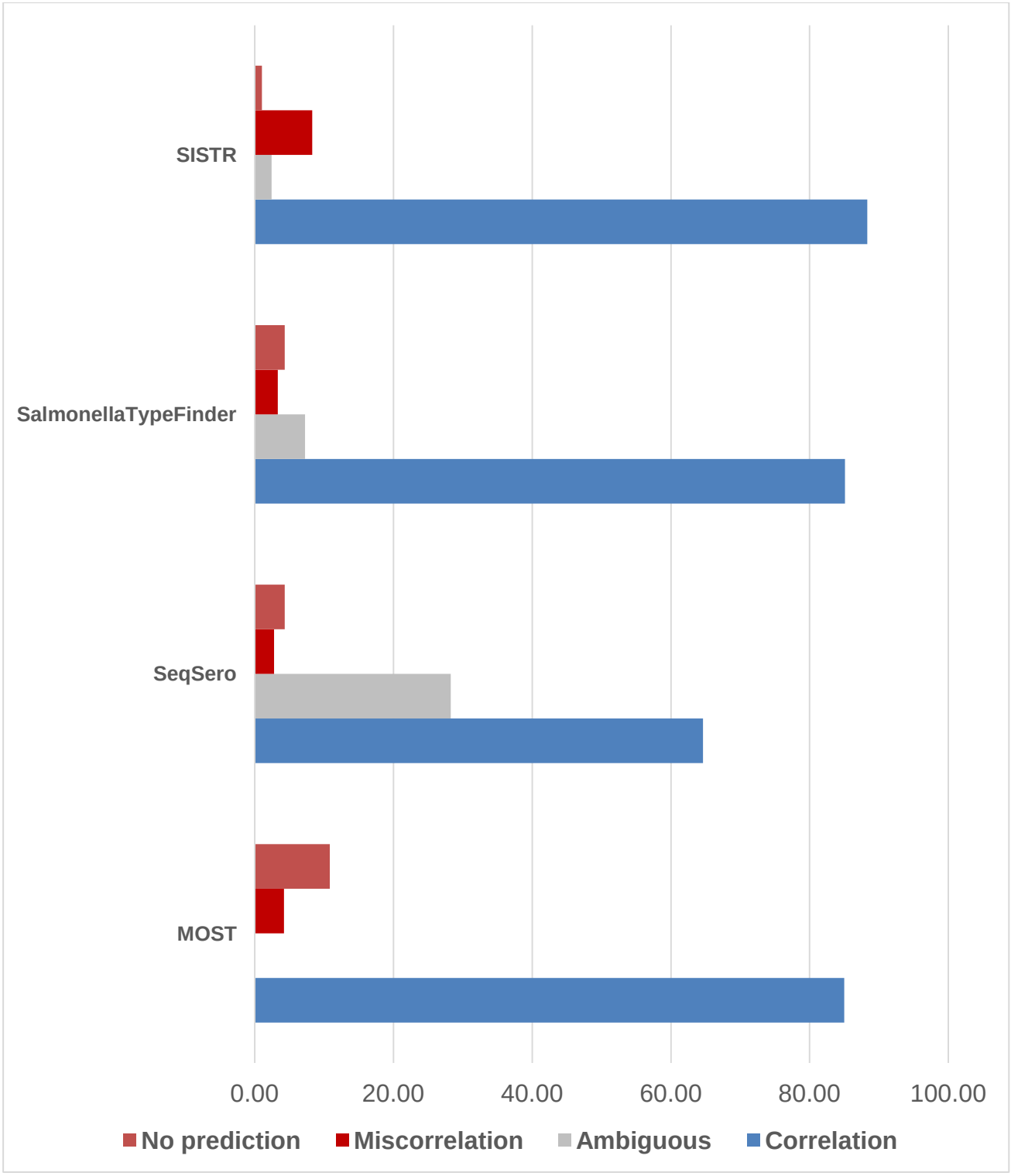
Y-axis represents number of isolates that serotype predictions yield no result.

Figure F.3: No prediction



Y-axis represents number of isolates that serotype predictions yield several possible serotypes, and the expected serotype is found among those.

Figure F.4: Ambiguous



X-axis represents percentage.

Figure F.5: Summary of correlation, miscorrelation, no prediction and ambiguous

Appendix G – Benchmarking of genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain – Method validation – Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures)

Report number	#3
Responsible	Eelco Franz (RIVM) and Pimlapas Leekitcharoenphon (DTU)
Other partners/institutions involved	Pimlapas Leekitcharoenphon (DTU), Liljana Petrovska (APHA), Kirsten Mooijman (RIVM), Eelco Franz (RIVM), Susanne Karlsmose Pedersen (DTU), Rene S. Hendriksen (DTU), Angela van Hoek (RIVM), Indra Bergval (RIVM), Rolf Sommer Kaas (DTU)
Benchmarking launched (date)	April 2017
Deliverable due (date)	June 2017

Purpose

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools for the *in silico* prediction of *Salmonella* serovars from raw whole genome sequencing data. The set-up of the interlaboratory (benchmarking) study complied with the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain – Method validation – Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures).

Participants

Participants in this benchmarking exercise were institutions from the ENGAGE network, including also participation from EFSA representatives, and from RIVM.

Thirteen sets of results were submitted from the following institutions:

APHA (United Kingdom), BfR (Germany), DTU (Denmark), EFSA (2 sets of results), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland), PHE (United Kingdom), RIVM (the Netherlands) (3 sets of results).

Participating institutes are identified by codes (1-13, see below) and each code is known only by the corresponding laboratory. The full list of laboratory codes is known only by the organizers (DTU).

Tools benchmarked

Benchmarking exercise component to determine species using the following tools and setup (each number refers to the corresponding participant and the (combination of) tools they used):

1. KmerFinder 2.1 (through Batch upload https://cge.cbs.dtu.dk/services/cge/index.php)
2. KmerFinder 2.0 (unix command line version, integrated in RIVM pipeline)
3. kmerid (PHE tool) tag version 2-1
4. CGE Tools (SPAdes 3.9, Assembler 1.2; SpeciesFinder 1.2; KmerFinder 2.0) (https://cge.cbs.dtu.dk/services/)
5. CGE KmerFinder 2.0. Scored method: winner takes it all. CGE SPADES 3.9 assembled sequences
6. SISTR (https://lfz.corefacility.ca/sistr-app/) 1.0.1; CGE KmerFinder 2.0 scored method: winner takes it all
7. CLC Genomics Workbench 9 & Species Finder 1.2 (https://cge.cbs.dtu.dk/services/SpeciesFinder/)
8. CLC Genomics Workbench 9 and SISTR app
9. CGE, blastn
10. Kraken version 0.10.5-beta, with MiniKraken DB. Options used: --fastq-input --gzip-compressed --quick --preload --paired
11. Blastn 2.3.0 -evalue 0.001 -outfmt "6 qseqid qlen sseqid sacc slen qstart qend sstart send evalue bitscore length pident mismatch gaps staxid sscinames" -perc_identity 95 -max_target_seqs 2 -qcov_hsp_perc 80 -db NT; for species confirmation: KmerFinder 2.0 (default options)
12. KmerFinder 2.0, scoring: Winner takes all, db: bacteria
13. KmerFinder 2.1 (BatchUpload of assembled data)

Benchmarking determining the *Salmonella* serovar genotypically using the following tools and setup:

1. SeqSero 1.0 (http://www.denglab.info/SeqSero) Reads paired-end
2. SISTR_cmd 0.3.6 (unix command line tool based on SISTR, integrated in our pipeline)
3. MOST (PHE tool) tag version 2-8, SeqSero (for antigenic formula) [-m 2 -b mem]
4. CGE Tools (SeqSero 1.2)
5. CGE SeqSero 1.2. We submitted raw sequences
6. SISTR (https://lfz.corefacility.ca/sistr-app/) v1.0.1
7. Seqsero 1.0 Genome Assembly and Species Finder 1.2
8. SISTR app
9. CGE (SeqSero, mlst)
10. SeqSero 1.0. Options used: -m2 (for pair-end reads), -b sam (for bwa samse/sampe)
11. SeqSero 1.0 -m2; for serotype confirmation: SISTR v0.3.4, --qc --no-cgmlst -f tab -o sistr-output.tab
12. SalmonellaTypeFinder 1.3
13. SeqSero 1.2 (paired end reads)

For further information on the serotyping tools, please see Appendix F – Benchmarking of genotypic *Salmonella* serotype prediction (general).

Genomes of bacterial species and *Salmonella* serovars

According to ISO/DIS 16140-6, the following number and type of strains have to be tested, per laboratory, in an interlaboratory study (ILS) when validating an alternative serotyping method (ISO/DIS 16140-6 includes protocols for validation of alternative confirmation and typing procedures i.e. including also serotyping) for *Salmonella*: 16 different strains from target serovars, 4 strains from non-target serovars within target subspecies and 4 strains from non-target genus. In this ILS, 27 genomes without any pre-assembly or trimming of the following strains were tested (the strains in this study were not part of ENGAGE project):

- 18 isolates of 6 target *Salmonella* serovars:
  - Enteritidis (n=3), Hadar (n=3), Infantis (n=3), monophasic Typhimurium (n=3), Typhimurium (n=3), Virchow (n=3).
- 5 non-target *Salmonella* serovars:
  - Derby, Dublin, Kentucky, Mbandaka, Stanley.
- 4 strains from the same family (Enterobacteriaceae) but non-target genus:
  - *Citrobacter freundii*, *Escherichia coli*, *Klebsiella pneumoniae*, *Shigella flexneri.*

The genomic quality based on number of reads, N50, number of contigs and total base pairs of each strain was assessed (Table G.6 – List of selected genomes) and they all were of good quality (genomic quality data can be found in the Supplementary Table 4 (Annex D)).

The National Institute for Public Health and the Environment (RIVM, the Netherlands), Centre for Zoonoses and Environmental Microbiology provided the genomes of six *Salmonella* genomes, serotyped by conventional methods, and one *E. coli* genome.

The Animal and Plant Health Agency (APHA, United Kingdom) provided seven *Salmonella* genomes, serotyped by conventional methods, one *Citrobacter freundii* genome and one *Klebsiella pneumoniae* genome, respectively.

The National Food Institute (DTU Food, Denmark) provided the 10 serotyped *Salmonella* genomes, serotyped by conventional methods and one *Shigella flexneri* genome.

All genomes were sequenced on either an Illumina MiSeq or Illumina HiSeq.

Overall results

The results were divided into the species and serovar predictions and correlated with the expected species and serovar (Tables G.1 and G.2 and Figures G.1 and G.2). The results that did not correlate with the expected result were further divided into predictions that give a different species and serovar than the expected (miscorrelation, Figures 1 and 2), predictions that yield no result (no prediction, Figures G.1 and G.2), and predictions that yield several possible serovars (ambiguous, Figures G.1 and G.2). Results are described into more detail in the Supplementary Table 4 in Annex D.

Table G.1: Correlation of *in silico* species prediction with conventional methods

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Correlation	25	25	26	24	24	25	25	26	26	25	26	25	25
No Correlation													
- Miscorrelation	2	2	1	2	1	1	2	1	1	2	1	1	2
- No prediction	0	0	0	1	2	1	0	0	0	0	0	1	0
- Ambiguous	0	0	0	0	0	0	0	0	0	0	0	0	0

Numbers represent the number of isolates. Total number of isolates is 27. Numbers in the header of the columns correspond to the listed participants for species prediction.

Table G.2: Correlation of *in silico* serovar prediction with conventional serotyping methods

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Correlation	20	22	22	17	19	22	16	22	20	20	22	22	20
No Correlation													
- Miscorrelation	1	1	1	0	1	1	1	1	1	1	1	1	1
- No prediction	0	0	0	4	1	0	6	0	0	0	0	0	0
- Ambiguous	2	0	0	2	2	0	0	0	2	2	0	0	2

Numbers represent the number of isolates. Total number of Salmonella isolates is 23. Numbers in the header of the columns correspond to the listed participants for serovar prediction.

Correlation for species prediction of all participants was more than 88%. Most of the tools failed to predict *Shigella flexneri* but identified *Shigella sonnei* instead. Almost all the tools predicted all *Salmonella enterica* correctly. The exception was ENGAGE-BM-16 which participants 4 and 5 did not predict correctly. Correlation of serovar prediction was between 74% and 96% The tools that resulted in a 96% correlation were SISTR (v1.0.1 and v0.3.6), SeqSero 1.0 (command line version), SalmonellaTypeFinder 1.3 and MOST. Most tools predicted *S.* Hadar (ENGAGE-BM-14) as *S.* Eko, meaning that these tools agreed upon the predicted serovar. Either conventional serotyping misclassified this isolate or the incorrect fastq files were added to the test panel. Many tools predicted *S.* Hadar (ENGAGE-BM-11 and ENGAGE-BM-13) ambiguously as *S.* Hadar/*S.* Istanbul. Colony form variation (the variable expression of minor antigens by different single-colony picks from the same strain) may occur with the expression of the O:6_1 antigen by some serogroup C2 serovars (Hendriksen et al., 2009; Popoff, 2001). SeqSero was the most used tool, however the correlation between the different versions (web-based or command line)/modes of input data (raw reads or assembled genomes) varied from 74.1% to 96.3%. This variation might be due to the choice of assembly tools, different options/parameters in web-based and command line version and to the operator. The second most used tool was SISTR that resulted in a 96.3% correlation.

Additionally, the results were evaluated following the data analysis and interpretation described in ISO/DIS 16140-6:2017. For this evaluation, the reference and alternative methods were compared for the target strains as well as for the non-target strains (inclusivity and exclusivity study, see Table G.3).

Table G.3: Comparison and interpretation of results between the reference and alternative methods for the inclusivity study (target strains) and for the exclusivity study (non-target strains)

Result of the (reference or alternative) method per strain		**Interpretation**
Reference confirmation procedure	**Alternative confirmation method**	**Alternative confirmation method compared to reference confirmation procedure**
+	+	PA
+	-	ND
-	+	PD
-	-	NA

PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement.

The results of the inclusivity and exclusivity analysis were compared to the acceptability limits (AL) indicated in ISO/DIS 16140-6:2017 (these acceptability limits are based on expert opinions), and are summarized in Table G.4 (species level) and Table G.5 (serovar level). For the evaluation at species level it was noticed that with two tools one *Salmonella* strain (BM-16) could not be identified and with three tools *Citrobacter freundii* (BM-06) was wrongly identified as *Salmonella* (Table G.4). For the evaluation at serovar level, the outcome '*S.* Hadar/*S.* Istanbul', instead of '*S.* Hadar' was still considered correct for reasons as described above. Additionally *S.* Hadar (ENGAGE-BM-14) was excluded from further analysis, because of inconsistent results between conventional and WGS serotyping. It was noticed that with three tools some *Salmonella* serovars could not be identified. These concerned 7 strains and in total 9 incidences (Table G.5).

Table G.4: Outcome inclusivity/exclusivity analysis at species level

	N	PA	ND	NA	PD	ND-PD	AL	ND+PD	AL
Inclusivity	299	297	2	0	0	2	3	2	5
Exclusivity	52	0	0	49	3	Not Applicable	Not Applicable	3	3

PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement; AL: Acceptability limits (in the ISO WG working on ISO 16140-6 it was agreed for the Exclusivity not to set targets for ND-PD).

Table G.5: Outcome inclusivity/exclusivity analysis at serovar level

	N	PA	ND	NA	PD	ND-PD	AL	ND+PD	AL
Inclusivity	221	212	9	0	0	9	3	9	5
Exclusivity	117	0	0	114	3	Not Applicable	Not Applicable	3	3

PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement; AL: Acceptability limits

Conclusions

The results of this benchmarking study demonstrate that serotyping using WGS data is a promising option. The tools predicting the *Salmonella* serovars in the most optimal way, in the current study, were, SISTR, SeqSero, SalmonellaTypeFinder followed by MOST, resulting in a 96.3% correlation with the conventional serotyping. This value was observed for MOST (only 1 participant used it), SISTR (3 participants used it), SeqSero (2 participants out of 8 who used this tool), and SalmonellaTypeFinder (only 1 participant used). The most optimal tool in this study based on unequal numbers of participants that used the tools. This was a limitation to evaluate the best tool in this study.

When analysing the data in accordance with ISO/DIS 16140-6:2017, the evaluation of results at species level showed to be within the acceptability limits, but at serovar level they exceeded these limits. This latter was mainly caused by the fact that in 9 incidences the *Salmonella* serovar of the target strains could not be identified. Testing non-target strains additional to target strains in such a study showed to be important as with 3 tools *Citrobacter* was incorrectly identified as *Salmonella*. The quality of the sequences and the choice of assembly tools and/or different options/parameter settings still need some attention when using WGS for serotyping *Salmonella* as participants who used different settings or assembly tools (also same tool using different online platforms), they got different serotyping results.

Additional notes

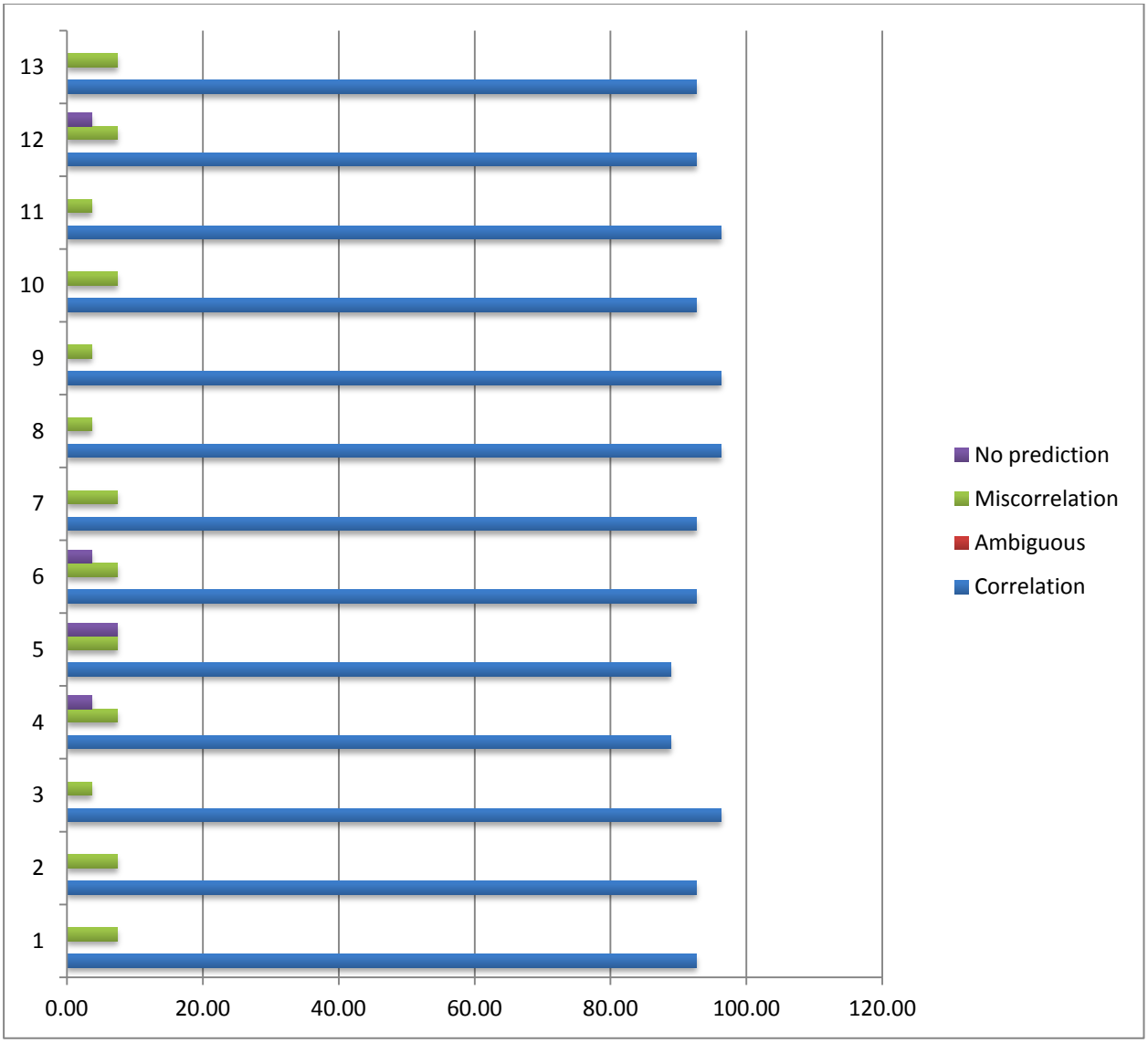
It is recommended to re-serotype, using the conventional serotyping, the isolates where the predictions from the tools disagree with the expected serovar.

References

ISO/DIS 16140-6:2017. Microbiology of the food chain — Method validation —Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures.

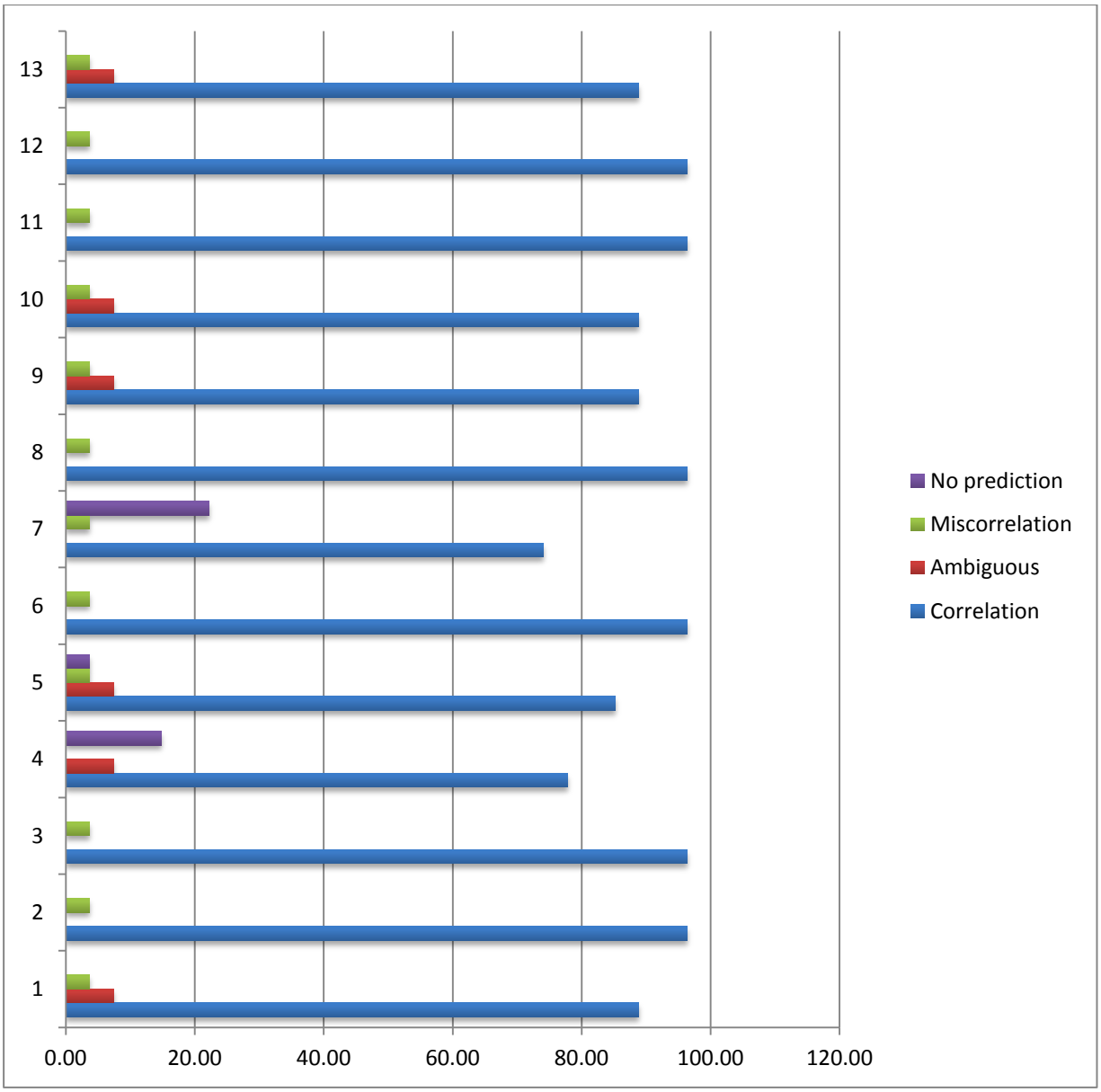
Hendriksen RS, Mikoleit M, Carlson VP, Karlsmose S, Vieira AR, Jensen AB, Seyfarth AM, DeLong SM, Weill FX, Lo Fo Wong DM, Angulo FJ, Wegener HC and Aarestrup FM, 2009. WHO Global Salm-Surv External Quality Assurance System for Serotyping of *Salmonella* Isolates from 2000 to 2007. Journal of Clinical Microbiology, 47(9), 2729-2736.

Popoff M Y, 2001. Guidelines for the preparation of *Salmonella* antisera, 6th ed. WHO Collaborating Centre for Reference and Research on *Salmonella.* Institut Pasteur, Paris, France.



X-axis represents percentage of correlation, ambiguous, miscorrelation and no prediction of species prediction.
Y-axis corresponds to the list of benchmark tools and participants for species prediction.

Figure G.1: Species prediction



X-axis represents percentage of correlation, ambiguous, miscorrelation and no prediction of Salmonella serotype prediction.
Y-axis corresponds to the list of benchmark tools and participants for species prediction.

Figure G.2: Serovar prediction

Table G.6: List of selected genomes

Target strains	RIVM	APHA	DTU
Typhimurium	ENGAGE-BM-20	ENGAGE-BM-03	ENGAGE-BM-01
monophasic Typhimurium	ENGAGE-BM-04	ENGAGE-BM-10	ENGAGE-BM-19
Enteritidis	ENGAGE-BM-21	ENGAGE-BM-25	ENGAGE-BM-09
Hadar	ENGAGE-BM-11		ENGAGE-BM-13, ENGAGE-BM-14
Infantis	ENGAGE-BM-15	ENGAGE-BM-22	ENGAGE-BM-16
Virchow	ENGAGE-BM-18	ENGAGE-BM-24	ENGAGE-BM-07
Non-target *Salmonella* serovars			
Dublin		ENGAGE-BM-26	
Stanley		ENGAGE-BM-27	
Derby			ENGAGE-BM-05
Kentucky			ENGAGE-BM-23
Mbandaka			ENGAGE-BM-08
Same family, but non-target genus			
Citrobacter freundii		ENGAGE-BM-06	
Escherichia coli	ENGAGE-BM-17		
Klebsiella pneumoniae		ENGAGE-BM-12	
Shigella flexneri			ENGAGE-BM-02

Selected (sequence data of) strains for interlaboratory study WGS serotyping *Salmonella*. Indicated are the strains selected per institute.

Appendix H – Benchmarking of genotypic detection of antimicrobial resistance (AMR) genes

Report number	#4
Responsible	Anthony Underwood (PHE)
Other partners/institutions involved	Pimlapas Leekitcharoenphon (DTU), Yue Tang (APHA), James Pettengill (FDA), Rolf Sommer Kaas (DTU), Kathie Grant (PHE), Liljana Petrovska (APHA), Rene S. Hendriksen (DTU), Susanne Karlsmose Pedersen (DTU)
Launch date	
Deliverable date	March 2017

Antimicrobial resistance (AMR) of foodborne pathogens is important for guiding treatment and surveillance of the antimicrobial resistance prevalence. Phenotypic susceptibility testing such as disk diffusion and Minimal Inhibitory Concentration (MIC) determination is a time-consuming and laborious process. Whole genome sequencing (WGS) offers an alternative to the phenotypic testing for determining the susceptibility to a range of antibiotics in a single test.

Purpose of the benchmarking exercise

The purpose of this study was to benchmark several of the currently available bioinformatics software tools for identification of AMR genes. A well-characterized set of food pathogen isolates (*Salmonella* and *E. coli*) that have been phenotypically tested for their susceptibility to several antimicrobials were compared to the genotypic profiles based on whole genome sequence data.

Benchmarked tools

The following tools with default parameters were assessed in the benchmarking exercise:

1) ResFinder 1.2 from DTU (available as command line and online tool)
   - BLAST-based detection of horizontally acquired genes and chromosomal point mutations (command line version)
2) KmerResistance 2.1 from DTU (available as command line and online tool)
   - Kmer-based detection of horizontally acquired genes (command line version used)
3) SRST2 v0.1.7 from http://katholt.github.io/srst2/ (available as command line only)
   - Mapping based detection of resistance genes
4) PHE Genefinder from PHE (available as command line only)
   - Mapping based detection of horizontally acquired genes and point mutations

Species/genomes included

Two datasets were collected for the purpose of this study.

The Animal and Plant Health Agency (APHA) collected 125 *Salmonella* isolates. Bacterial DNA was extracted using the MagNA Pure LC DNA Isolation Kit III (Roche) according to manufacturer's instructions and sequencing libraries were prepared using the NexteraXT sample preparation method

for sequencing on the Illumina HiSeq platform with paired-end 2x125bp reads (http://www.illumina.com).

The National Food Institute at DTU collected 164 *E.coli* isolates. Genomic DNA was extracted using an Invitrogen Easy-DNA™ Kit (Invitrogen, Carlsbad, CA, USA) and DNA concentrations were determined using the Qubit dsDNA BR assay kit (Invitrogen). The genomic DNA was prepared for Illumina pair-end sequencing using the Illumina (Illumina, Inc., San Diego, CA) NexteraXT® Guide 150319425031942 following the protocol revision C (http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html). A sample of the pooled NexteraXT Libraries was loaded onto a Illumina HiSeq reagent cartridge using HiSeq Reagent Kit v2. The libraries were sequenced using an Illumina Hiseq platform.

The final datasets consisted of 289 isolates. Raw reads were trimmed using bbduk2 (part of the suit bbtools version 36.49) with the following cut-off; 1) length of read >= 50 bp, 2) Phred score per base >= 20. De novo assembly was performed using SPAdes with minimum Kmer coverage at 2 and minimum contig size at 500 bp.

For dataset from APHA, the antimicrobial susceptibility testing of the 125 *Salmonella* isolates was performed and interpreted using the Kirby-Bauer disk diffusion method, on Isosensitest Agar (Oxoid) as described by the British Society for Antimicrobial Chemotherapy (BSAC). The following antimicrobials were included in the testing with the listed disc concentrations (µg per ml): nalidixic acid (30); tetracycline (10); neomycin (10); ampicillin (10); furazolidone (15); ceftazidime (30); sulphamethoxazole/trimethoprim (25); chloramphenicol (30); amikacin (30); amoxicillin/clavulanic acid (30); gentamicin (10); streptomycin (10); sulphonamide compounds (300); cefotaxime (30); apramycin (15); ciprofloxacin (1).

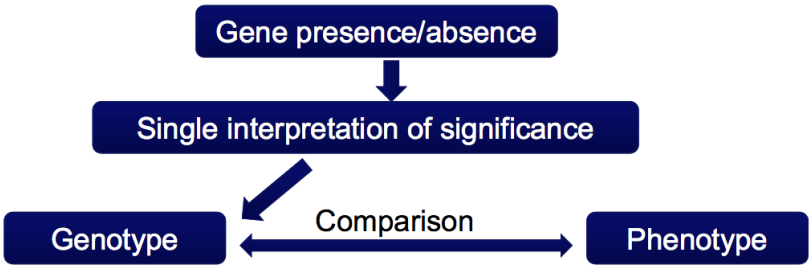
Minimum inhibitory concentration (MIC) determination was performed at DTU on the 164 *E.coli* isolates using commercially prepared dehydrated panels, EUVSEC and EUVSEC2 (Sensititre; TREK Diagnostic Systems Ltd., East Grinstead, England). EUCAST epidemiological cut-off values were used as interpretative criteria to determine the phenotypic resistance (http://www.eucast.org). Quality control was performed by using reference strain *E. coli* ATCC 25922 according to CLSI guidelines.

Method

Four tools were benchmarked in this study: ResFinder (Zankari et al., 2012), KmerResistance (Clausen et al., 2016), SRST2 (Inouye et al., 2014), PHE Genefinder.

Availability of tools:
ResFinder: https://cge.cbs.dtu.dk/services/ResFinder/
KmerResistance: https://cge.cbs.dtu.dk/services/KmerResistance/
SRST2: https://katholt.github.io/srst2/
PHE Genefinder (in-house software, not publicly available)

The genotypic testing was performed independently by the different collaborating partners and results were afterwards compared. Thus, DTU tested the ResFinder tool, PHE tested the KmerResistance tool and the Genefinder, and APHA tested the SRST2 tool. The phenotypic susceptibility data were used as proxy for the true result. The genotypic results were compared to the phenotypic susceptibility data and the performance of the tools was assessed by calculating the specificity, sensitivity, accuracy and the Matthew's Correlation Coefficient (MCC). Additional statistical tests of agreement were also applied based on learning from the serotype benchmarking exercise. Given that resistance to different classes of antibiotics is conferred by different genes, it was decided to break down the results by antibiotics since some of the software tools perform better at calling a profile for different classes.



Overall results

The results for specificity, sensitivity, accuracy and MCC for all antibiotic classes are presented in Table H.1 and Table H.2 and the accuracy in predicting different classes of antibiotic in Figure H.1, Figure H.2 and Supplementary Table 5 (Annex E).

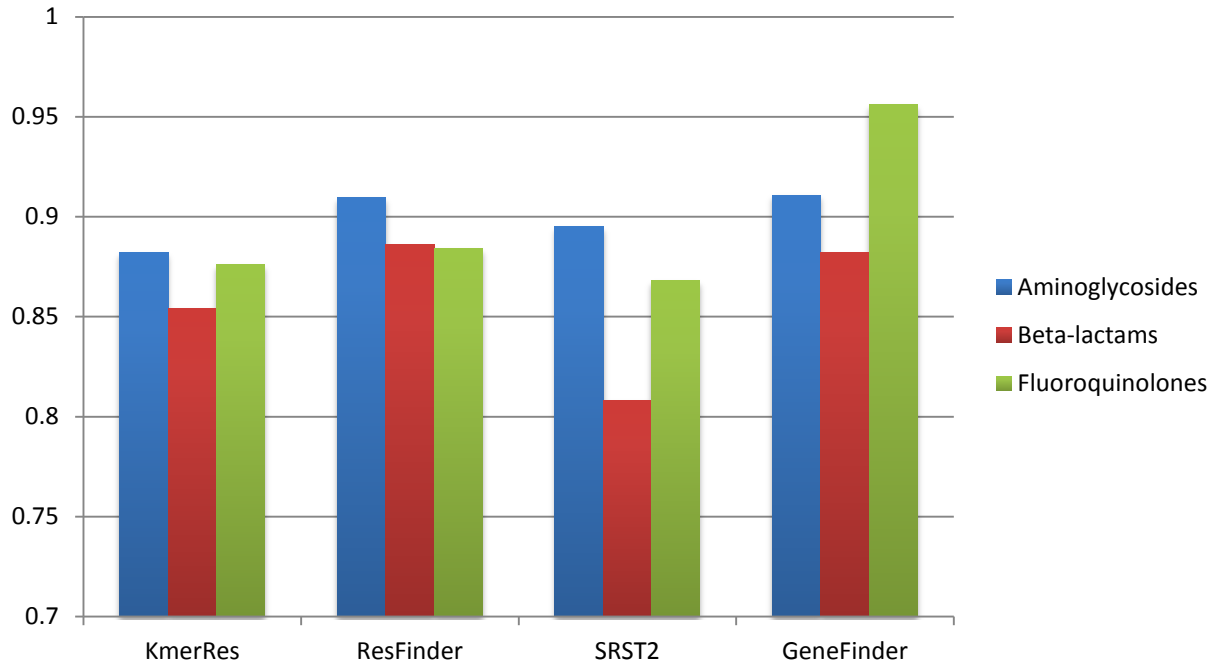
All tools tested provided an approximate accuracy of around 90% when testing the *Salmonella* genomes (Table H.1). All tested tools achieved an overall lower accuracy, between 80-82%, when testing the *E. coli* dataset (Table H.2). The accuracy in predicting resistance in *E. coli* for ß-lactams and fluoroquinolones using all tools was low, ranging between 55% - 58% and 82% - 84%, respectively (Figure H.2 and Supplementary Table 5 in Annex E).

Table H.1: Results from *Salmonella* dataset for all antibiotic classes

Software	Specificity	Sensitivity	Accuracy	MCC
KmerResistance	0.95	0.74	0.86	0.72
ResFinder	0.95	0.83	0.90	0.79
SRST2	0.93	0.80	0.87	0.74
PHE GeneFinder	0.97	0.83	0.90	0.81

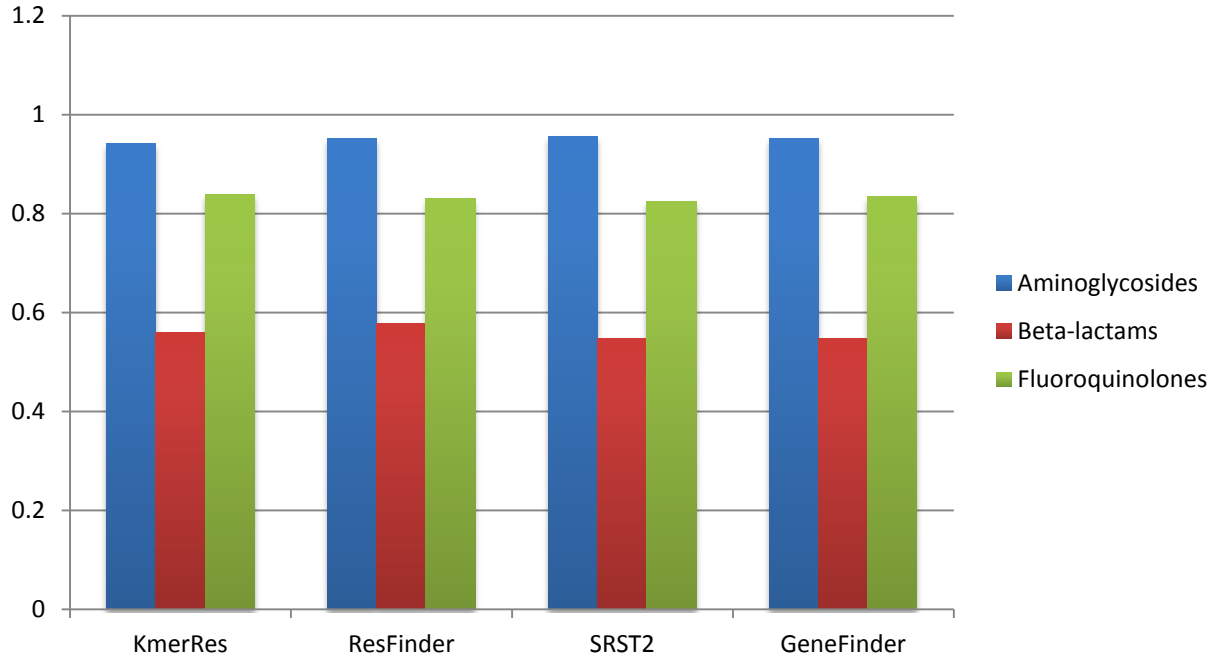
Table H.2: Results from *E. coli* dataset for all antibiotic classes

Software	Specificity	Sensitivity	Accuracy	MCC
KmerResistance	0.91	0.46	0.80	0.41
ResFinder	0.89	0.60	0.82	0.51
SRST2	0.89	0.57	0.81	0.48
PHE GeneFinder	0.90	0.53	0.81	0.47



Y-axis represents accuracy ratio expressed as a fraction of 1.

Figure H.1: Accuracy obtained by the benchmarked tools for three antimicrobial classes for the tested *Salmonella* dataset



Y-axis represents accuracy ratio expressed as a fraction of 1.

Figure H.2: Accuracy obtained by the benchmarked tools for three antimicrobial classes for the tested *E.coli* dataset

Conclusion

The tools providing the highest degrees of specificity, sensitivity, MCC and accuracy in *Salmonella* data were the ResFinder 1.2 and PHE GeneFinder tools (no version available; tests performed on 01.02.2017). ResFinder also provided the highest accuracy and MCC in predicting resistance in the *E. coli* genomes, while GeneFinder provided the highest MCC in predicting resistance in the *Salmonella* genomes.

All tools revealed an approximate 90% correlation with the phenotypic susceptibility testing for *Salmonella*. Only the PHE Genefinder predicted resistance to fluoroquinolones based on chromosomal point mutation and hereby performed with a higher accuracy than other tools for fluoroquinolone resistance (Figure H.1).

All tools performed with a lower accuracy when testing *E. coli*. A very low accuracy was achieved in profiling β-lactam (Figure H.2). This could be due to the possible bias in the dataset that included a number of *E. coli* containing upregulated chromosomal *ampC* mutations (mediating β-lactam resistance) which none of the tools could predict. By including the methods to detect *ampC* mutations and other chromosomal point mutations, the concordance for β-lactam and fluoroquinolone resistance can be increased. Therefore, *ampC* mutations and other chromosomal point mutations will soon be included in a new version of ResFinder.

The results of this benchmarking study showed that predicting antimicrobial resistance using WGS is a feasible and realistic alternative to phenotypic susceptibility testing. In addition, for the *Salmonella* and the *E.coli* datasets, different criteria were applied for the definition of phenotypic resistance, as this can also influence the results. The comparability of the phenotypic results should be taken into account because phenotypic criteria for defining resistance and susceptible were different. This might affect the results on the correlations between phenotypes and genotypes.

The miscorrelation rate (cases where the tools predicted a different antimicrobial profile than the expected) were 10-14% in the *Salmonella* dataset. These miscorrelations are suspected to be caused by mistakes in the phenotypic susceptibility testing.

Additional notes

It is recommended to retest the phenotypic susceptibility for the isolates showing discordant results with the genotypic prediction tools and the expected genotypic resistance profile derived from phenotypic susceptibility testing. This is especially important for the isolates where all tools showed identical miscorrelations. Additionally, the sequencing quality also influences the performance of the tools.

References

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM and Larsen MV, 2015. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotheraphy, 67(11):2640-2644. doi: 10.1093/jac/dks261

Clausen PT, Zankari E, Aarestrup FM and Lund O, 2016. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. Journal of Antimicrobial Chemotheraphy, 71, 2484–2488. doi: 10.1093/jac/dkw184

Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J and Holt KE, 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine, 20;6(11):90. doi: 10.1186/s13073-014-0090-6.

Appendix I – Benchmarking for *Salmonella Enteritidis* phylogeny

Report number	#5
Responsible	Anaïs Painset (PHE) and Anthony Underwood (PHE)
Other partners/institutions involved	APHA (United Kingdom), BfR (Germany), EFSA (Italy), DTU (Denmark), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland)
Benchmarking launched (date)	September 2017
Deliverable due (date)	October 2017

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools both to detect variants and to build a phylogeny based on the variants alignment detected for *Salmonella* Enteritidis isolates. With the use of Whole Genome Sequencing, phylogeny is used as a method to characterize microorganisms in outbreak investigations and for surveillance of isolates that are genetically related.

Participants

Participants in this benchmarking were institutions from the ENGAGE network.

Twelve sets of results (phylogenies) were submitted from the following institutions:

APHA (United Kingdom), BfR (Germany), DTU (Denmark), EFSA (Italy), IZSLT (Italy), IZSVe (Italy) (3 phylogenies), NIPH-NIH (Poland), NVRI (Poland) (2 phylogenies), PHE (United Kingdom).

Results from participating institutes are identified by codes (1-12 see below) and each code is known only by the corresponding laboratory. The full list of laboratory codes is known only by the organizers (PHE).

Tools benchmarked

Benchmarking by variants calling and generating SNPs alignment using the following tools and setup:

1. Snippy v3.0 [default setting: min depth 10, 90% difference from ref]
2. BioNumerics 7.6 (- Mapping /SNP Filtering (relative coverage: total: 5, forward: 1, reverse: 1, unreliable bases, ambiguous bases, gaps, non-informative SNPs))
3. CGE Tools (command line version) – CSIPhylogeny v1.4
4. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
5. CGE Tools CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
6. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
7. CGE Tools – CSIPhylogeny v1.4 online version default parameters and reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)

8. CGE Tools – CSIPhylogeny v1.4 online version default parameters and reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
9. CGE Tools – CSIPhylogeny v1.4 online version, reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
10. Custom pipeline
    • Trimmomatic v.0.36 and Nextera-PE adapters to trim the reads. Following parameters were set:
    ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

    • BWA MEM v.0.7.13 with default settings for mapping paired and unpaired reads (after trimming)

    • Freebayes v.1.1.0 (d784cf8) with "--ploidy 1" for joint variant calling on all samples

    • R package VariantAnnotation v1.22.3 to filter variants: variant calls with genotype-likelihood (GL) > -30 (likelihood > 10e-3) were set to unknown genotype.

    • VCF-kit v.0.1.2 "pheno fasta" and "pheno tree nj" to generate alignment and newick tree
11. PHEnix 1.2 (BWA mapping + GATK variant calling) + SnapperDB 0.2.4 (get the snps A:80)
12. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)

Benchmarking by building trees using the following tools and setup:

14. RAxML v.8.2.9
15. Bionumerics v.7.6: Neighbor joining tree
16. CGE Tools – CSIPhylogeny 1.4 command line (FastTree built-in)
17. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
18. MEGA-CC 7 Minimum Evolution Methods
19. MEGA-CC 7 Maximum Parsimony
20. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
21. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
22. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
23. VCF-kit v0.1.2 with pheno tree nj
24. RAxML v8.2.8-multithread (-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT)
25. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)

Species/genomes included

Public Heath England selected and provided genomes from *Salmonella enterica* serotype Enteriditis and part of the same eburst group EGB4. The genomes have been selected because they were part of an outbreak investigated by PHE. This outbreak has been well-studied (Dallman et al., 2016) and epidemiological information support the phylogeny associated with the selection.

Thirty genomes represented by sets of fastq (paired) were included in the data set (Annex F). All genomes originated from sequencing using an Illumina HiSeq. Fastq were trimmed using Trimmomatic 0.32 with the following options: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:8:true LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50, then the quality of the sequencing was assessed by running FastQC 0.11.3. The trimmed and quality assessed reads were used for the analysis (Supplementary Table 6 in Annex F).

The gold standard phylogeny use to perform comparisons was constructed following the methods employed by Centre 11. Tools used to build the gold standard phylogenies are PHEnix 1.2 for variants calling and filtering, follow by SnapperDB 0.2.4 to extract relevant SNPs and RAxML 8.2.8 to build the phylogeny. In this benchmarking, gold standard will be the phylogeny build following Centre 11 tools/methods.

Overall results

The results were compared using two main approaches:

1. Alignment and distance matrix comparison
2. Topology of the tree: global topology, Robinson-Fould symmetric difference and percentage of edge similarity (number of branches in one tree that are present in another)

Alignment and distance matrix

All the participants were required to provide a fasta alignment of the SNPs detected by the method they employed to generate the phylogeny. To ensure consistent comparison of the alignments, we generated the distance matrices from the alignment using an in-house script and build the graphic with an in-house R script.

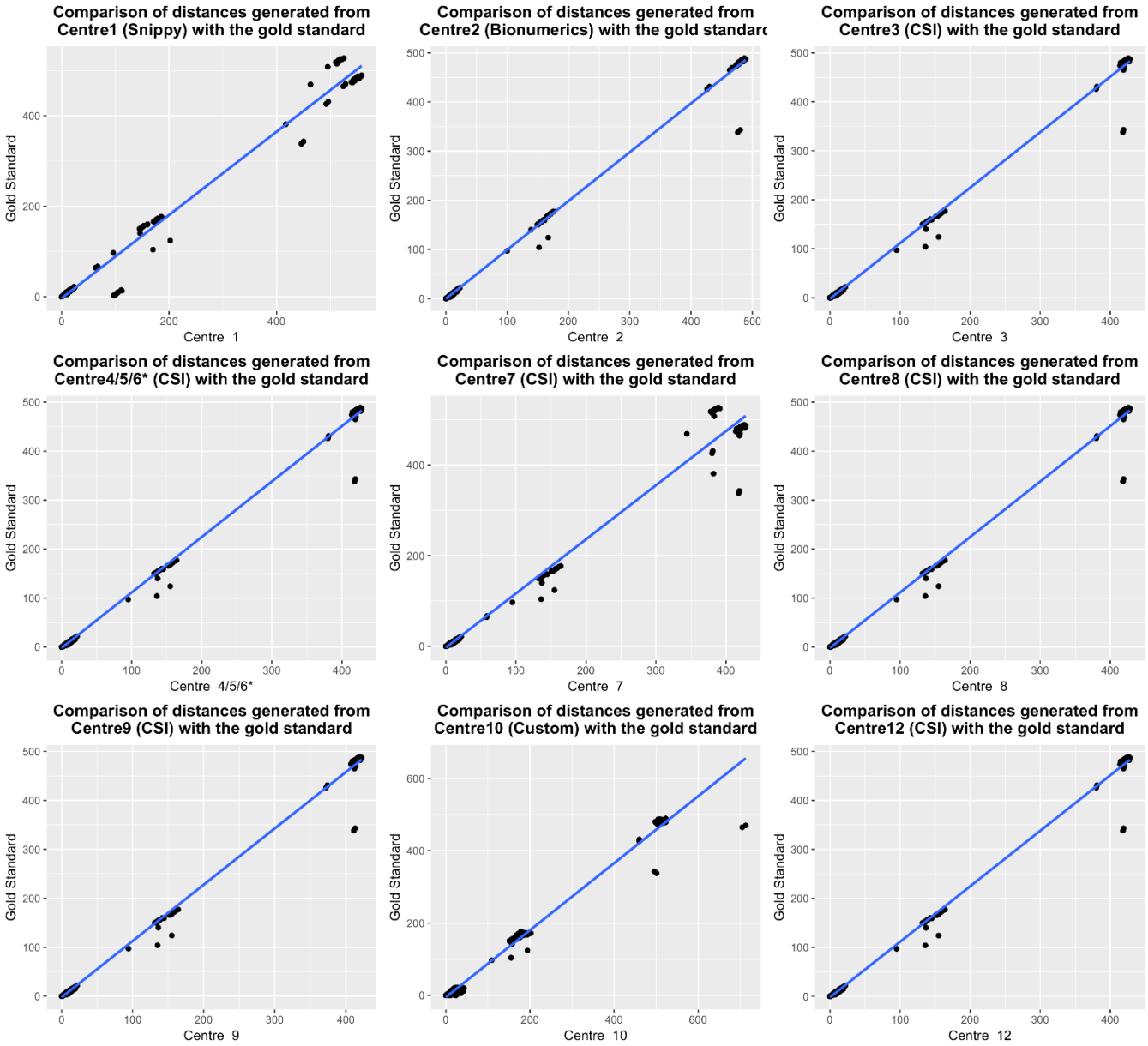
Eight out of the twelve set of results provided by the partners were generated by using the CGE CSI Phylogeny tools. We decided to regroup results in Table I.1 where the parameters were the same.

Table I.1: Alignment and statistic metrics

Results	1	2	3/4/5/6/12	7/8	9	10	11
Alignment length	779	698	633	644	636	1465	786
Min distance matrix	0	0	0	0	0	0	0
Max distance matrix	558	489	428	428	422	713	527
Reference	+	-	-	+	+	-	+

Columns numbers correspond to the results submit by the partners. The list of benchmarking tools and participants for variants calling.

The longest the alignment is the more SNPs have been detected in the dataset. Min and max distance matrix represent the number of SNP different between strains in the dataset. Strains supposedly part of an outbreak or closely related are expected to have a low number of SNP difference. The minimum distance captures the minimum number of SNPs between two strains in the dataset i.e. the two closest strains in the dataset. The maximum distance reflects the maximum number of SNPs between two strains in the dataset i.e. the more distant strain in the dataset.



Method use by centre 11 was used as the gold standard and is not represented on the comparisons.
* Phylogenies 4/5/6 were based on the same alignment, therefore only one graph can be produce.
Centre numbers correspond to the list of benchmarking tools and participants for variants calling.

Figure I.1. Comparisons of distances generated from centre with gold standard.

Topology of the tree

All the phylogenies are presented in the Figures I.4 - I.16. They are labelled according to row number on the following table. The phylogenetic distance metrics were generated by using the ete toolktit (http://etetoolkit.org/) ete3 v.3.0.0 with his module compare and the additional phangorn R package v2.0.0.

Table I.2: Phylogenetic distance metrics

	1	2	3	4	5	6	7	8	9	10	11	12
E.size	31	30	30	30	30	30	31	31	31	30	31	30
Ref	+	-	-	-	-	-	+	+	+	-	+	-
nRF	0.46	0.32	0.2	0.2	0.37	0.37	0.19	0.19	0.19	0.78	0	0.2
RF	26	13	10	10	20	20	10	10	10	42	0	10
maxRF	56	41	50	50	54	54	52	52	52	54	56	50
src-br+	0.78	1	0.94	0.94	0.82	0.82	0.94	0.94	0.94	0.62	1	0.94
ref-br+	0.78	0.77	0.88	0.88	0.82	0.82	0.88	0.88	0.88	0.62	1	0.88
KF.dist	0.198	356.692	0.073	0.073	313.742	-	0.151	0.151	0.166	0.395	0	0.073

Columns numbers correspond to the list of benchmarking tools and participants for the tree building.
Additional notes: meaning of the metrics (ete-compare):

E.SIZE: effective size of the dataset used to calculate metrics

nRF: Normalized Robinson-Foulds distance (RF/maxRF)

RF: Robinson-Foulds symmetric distance

maxRF maximum Robinson-Foulds value for this comparison

%src_br (percent source branch): frequency of edges in target tree found in the reference (1.00 = 100% of branches are found)

%ref_br (percent reference branch): frequency of edges in the reference tree found in target (1.00 = 100% of branches are found)

KF.dist (Kuhner-Felsenstein distance): branch score distance (Kuhner & Felsenstein 1994) [compute with Phargorn].

The closer the normalized Robinson-Foulds (nRF) value is to 0, the better the match of the topology to the 'gold standard' phylogeny. As we can see most of the trees are close to the reference one. One tree (Centre 10) is significantly different in terms of topology compared to the gold standard.

The KF distance (KF.dist) measures the difference in term of branch length. As we can see most of the trees have really similar branch length. Tree of centre 2 and tree of centre 5 are not using the SNPs for the alignment as a branch length and this would explain why the difference in term of branch length is really high.

The tree of centre 6 does not provide branch length in the newick file and therefore was exclude from this metric (Table I.2).

Table I.3: Clade retrieval from gold standard compared to others methods

	1	2	3	4	5	6	7	8	9	10	11	12
Reference	+	-	-	-	-	-	+	+	+	-	+	-
Outliers n=5	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade I n=3	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade II n=8	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade III n=14	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y

Columns numbers correspond to the list of benchmarking tools and participants for the tree building.
n = number of isolates per clade, N if all isolates from the gold standard clade are not retrieved on the same clade in the phylogeny, Y if all isolates from the gold standard clade are retrieve on the same clade. +/- indicated presence/absence of the reference in the phylogeny.

This is a topological assumption: isolates from a clade are considered correct if they are on the same monophyletic branch. The three clades should be separated from the outliers by a long branch. The assumption is that isolates group in clades accordingly to the gold standard (Table I.3).

The following tanglegrams illustrate the difference/similarity between the gold standard and the phylogeny where we found clade discrepancies. The lines in the middle reflect inversions in the position of isolates between the two phylogenies; it is used to illustrate the most different trees in terms of clade retrieval compare to the gold standard.

Figure I.2. Tanglegram of the gold standard (left) versus the most different topology produce by Centre 10

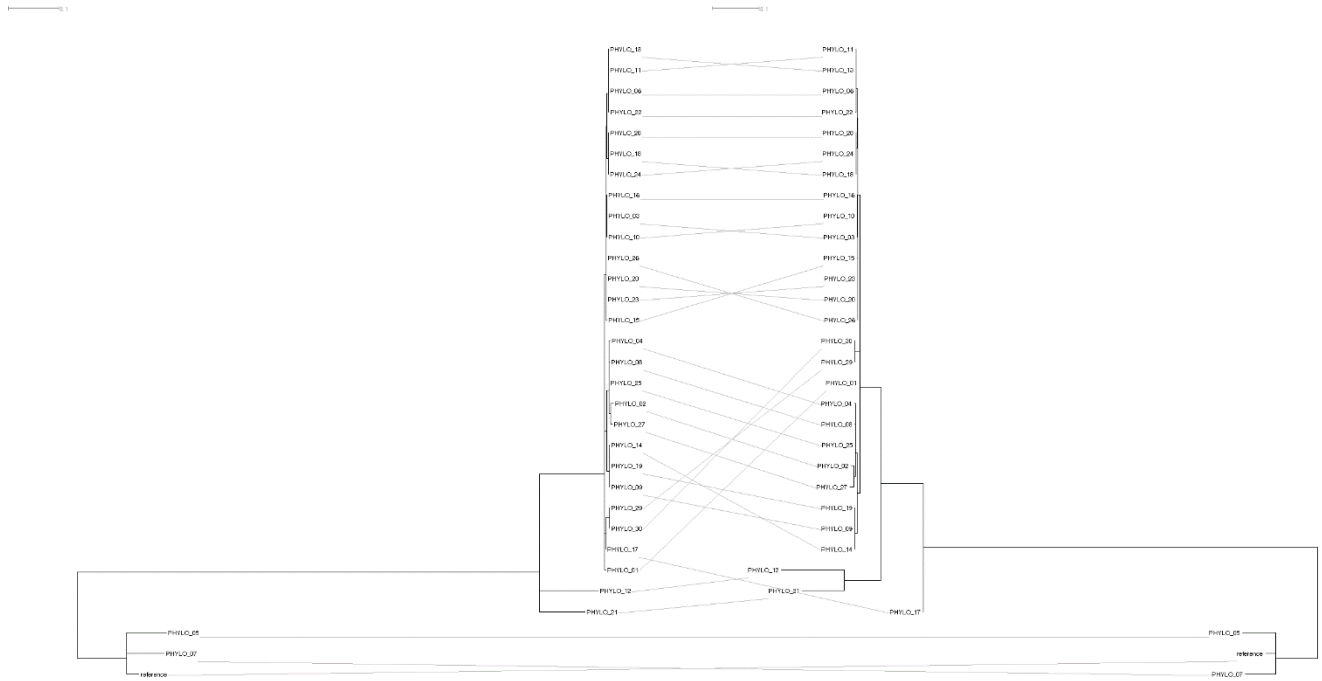


Figure I.3. Tanglegram of the gold standard (left) versus the topology produce by Centre 1

Conclusion

The methods used to generate the SNP alignment by the different partners showed similar results except for three Centres, Centre 1 (Snippy), Centre 7 (CSI without heterozygote removal) and Centre 10 (Custom pipeline) where the comparisons of the distance matrix show discrepancies between those isolates that are distantly related. Also, the topology produced shows great similarity, the number of SNPs difference between isolates can vary based on the tools and parameters.

The scores based on the topology demonstrate that most of the methods tested are able to retrieve the topology derived from the gold standard. Only one method seems to give a markedly different topology (Centre 10, custom pipeline).

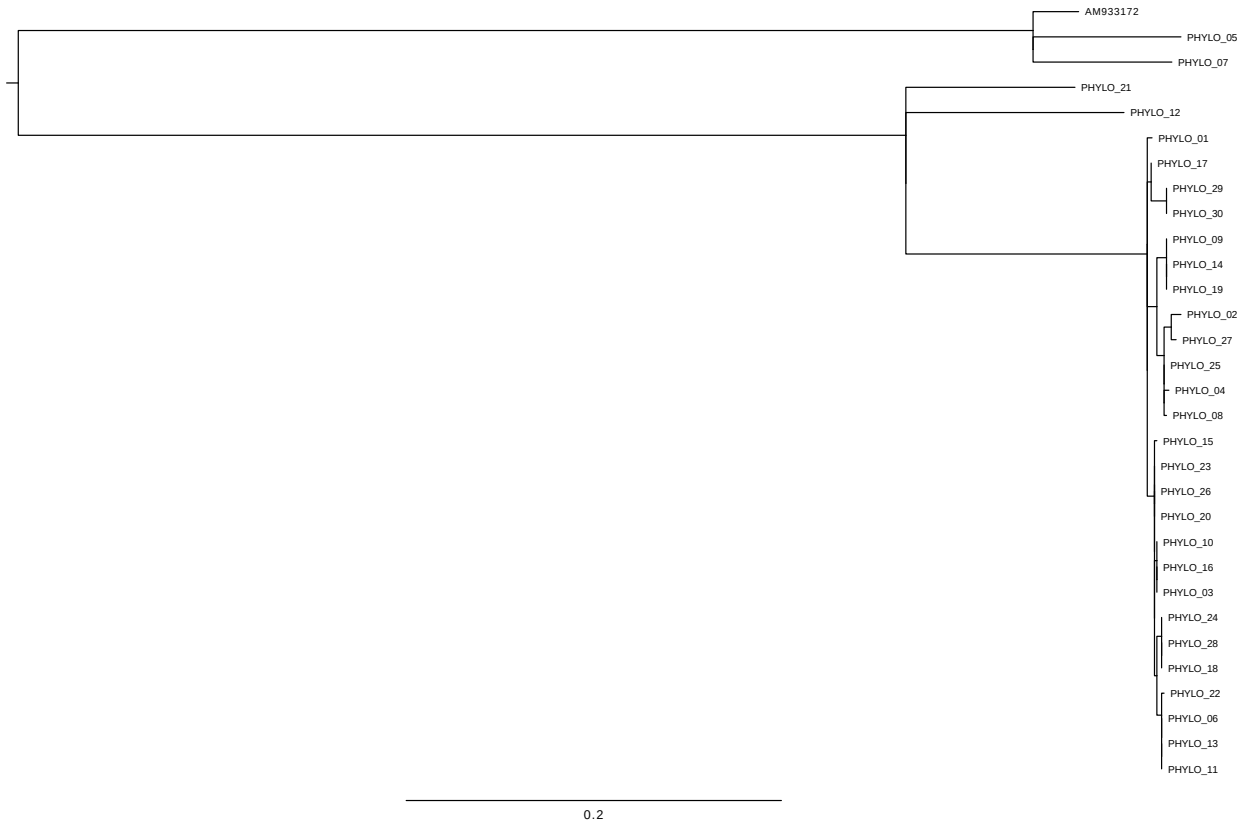
During this benchmarking we have identified that a key point in building a phylogeny based on the SNP differences between isolates is the detection and filtering of the SNPs. Based on this benchmarking we can recommend a minimum depth coverage for the SNPs detection > 10, a minimum mapping read quality of 30, and 90% consensus for the reads mapped at a position that differs from the reference.

The best tools to build tree from an alignment are maximum likelihood methods. Topology obtained using these methods produce trees with the best correlation between gold standard and the obtained phylogeny.

Reference

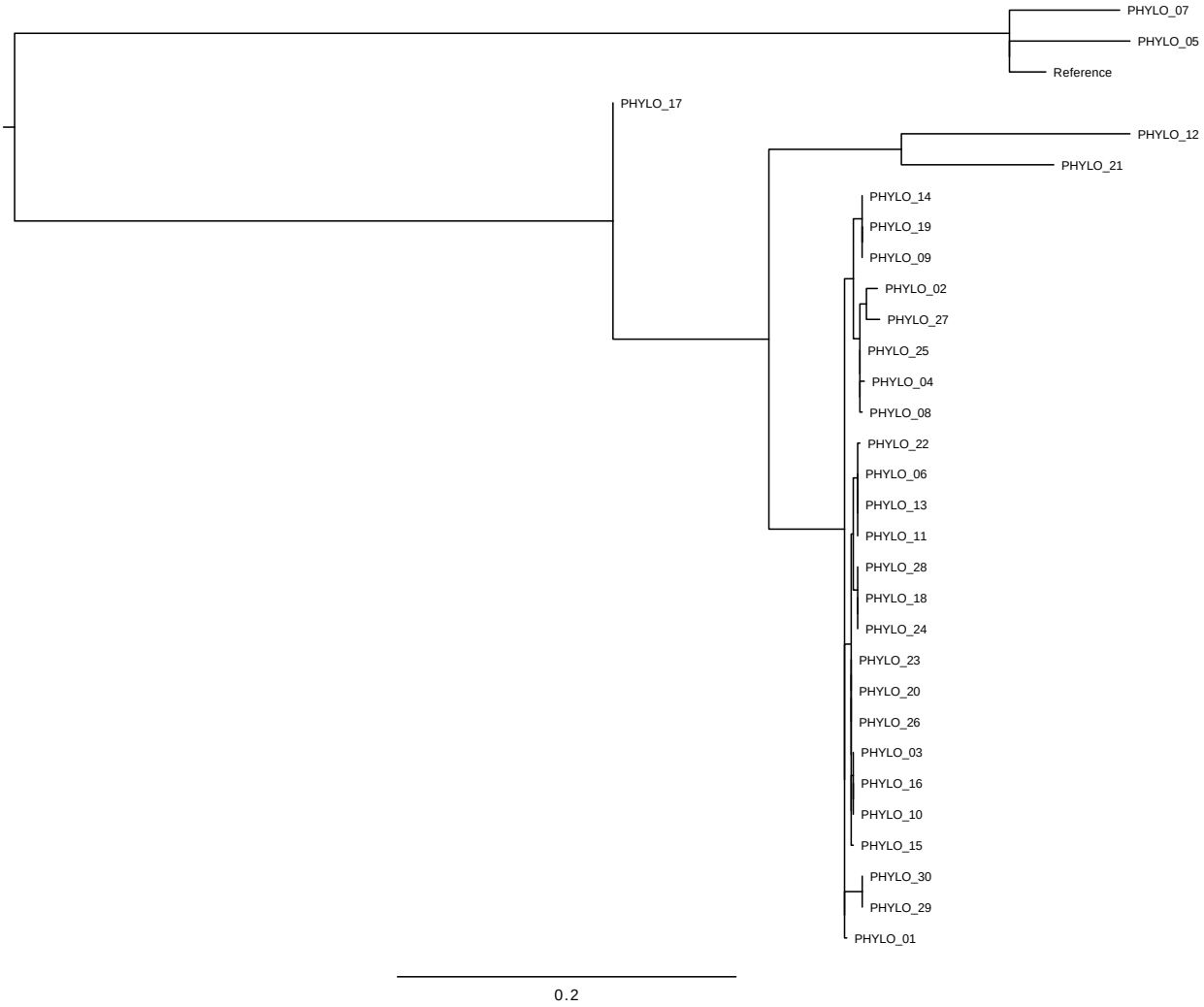
Dallman T, Inns T, Jombart T, Ashton P, Loman N, Chatt C, Messelhaeusser U, Rabsch W, Simon S, Nikisins S, Bernard H, le Hello S, Jourdan da-Silva N, Kornschober C, Mossong J, Hawkey P, de Pinna E, Grant K and Cleary P, 2016. Phylogenetic structure of European *Salmonella Enteritidis* outbreak correlates with national and international egg distribution network. Microbial Genomics, 2(8):e000070. doi: 10.1099/mgen.0.000070

Additional Figures



Scale represents the branch length stipulated into the newick file.

Figure I.4: Gold standard phylogeny with reference

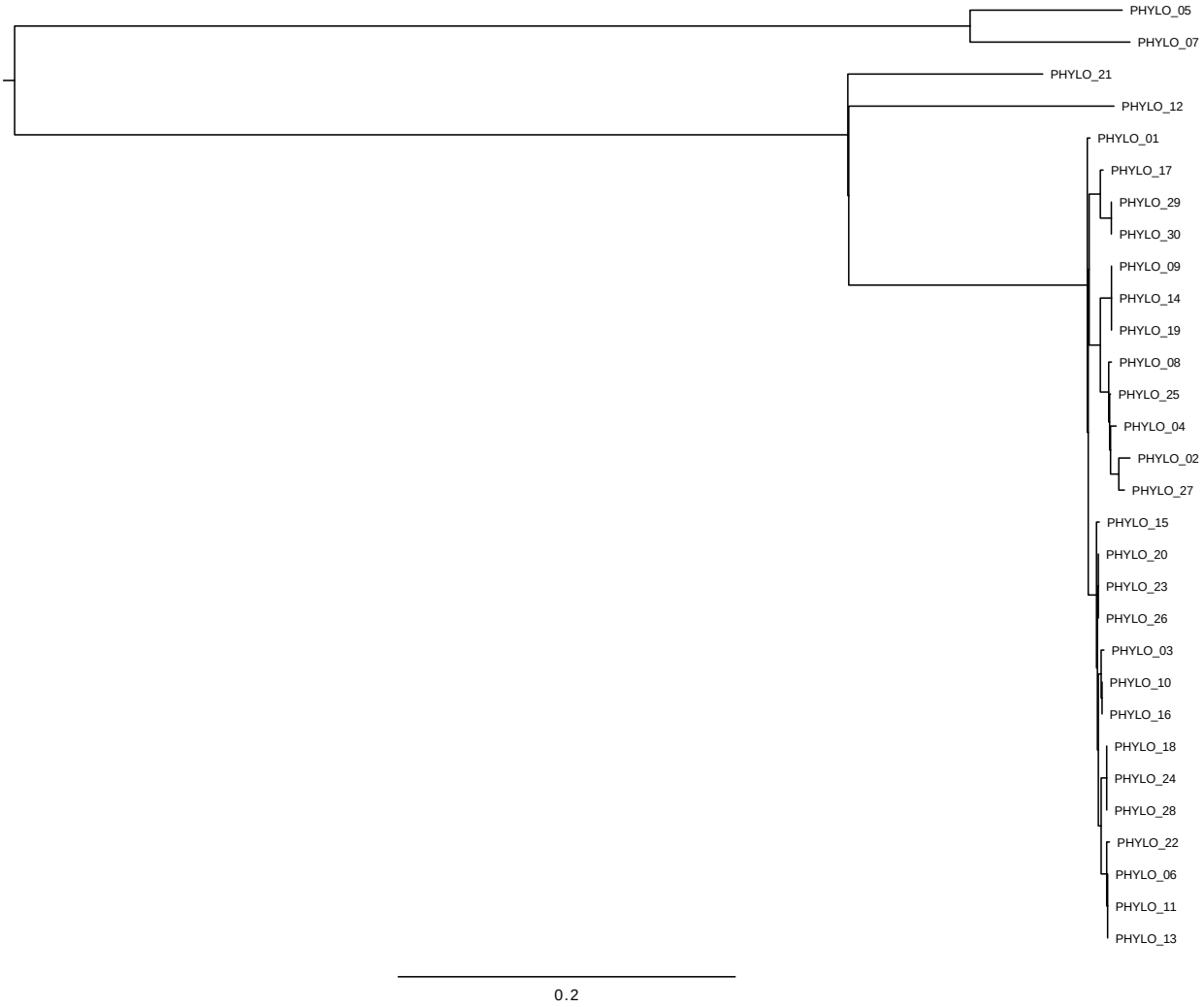


Scale represents the branch length stipulated into the newick file.

Figure I.5: Phylogeny Centre 1 obtained with Snippy tool and RAxML

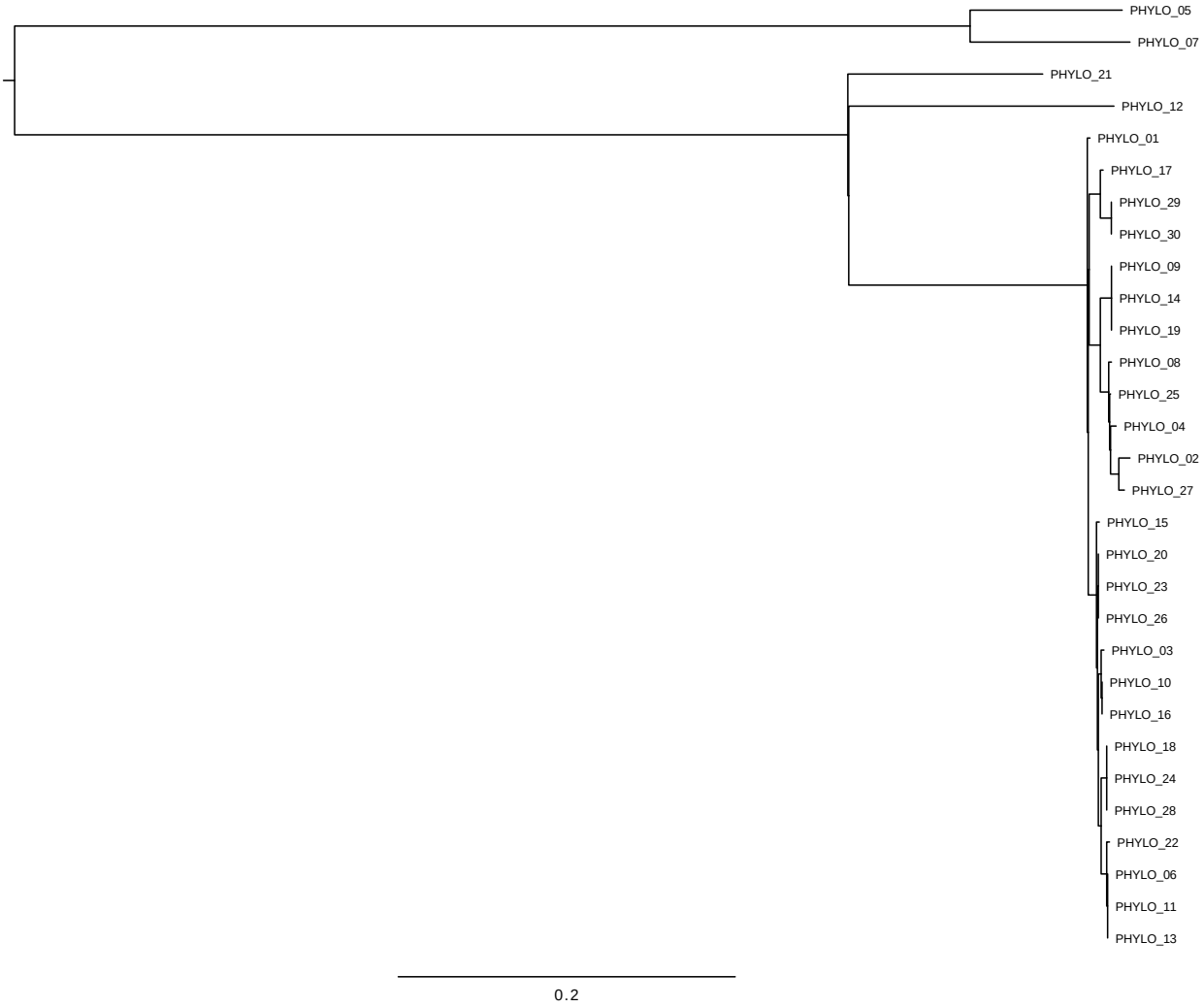
Scale represents the branch length stipulated into the newick file.

Figure I.6: Phylogeny Centre 2 obtained with BioNumerics and a Neighbor joining tree reconstruction



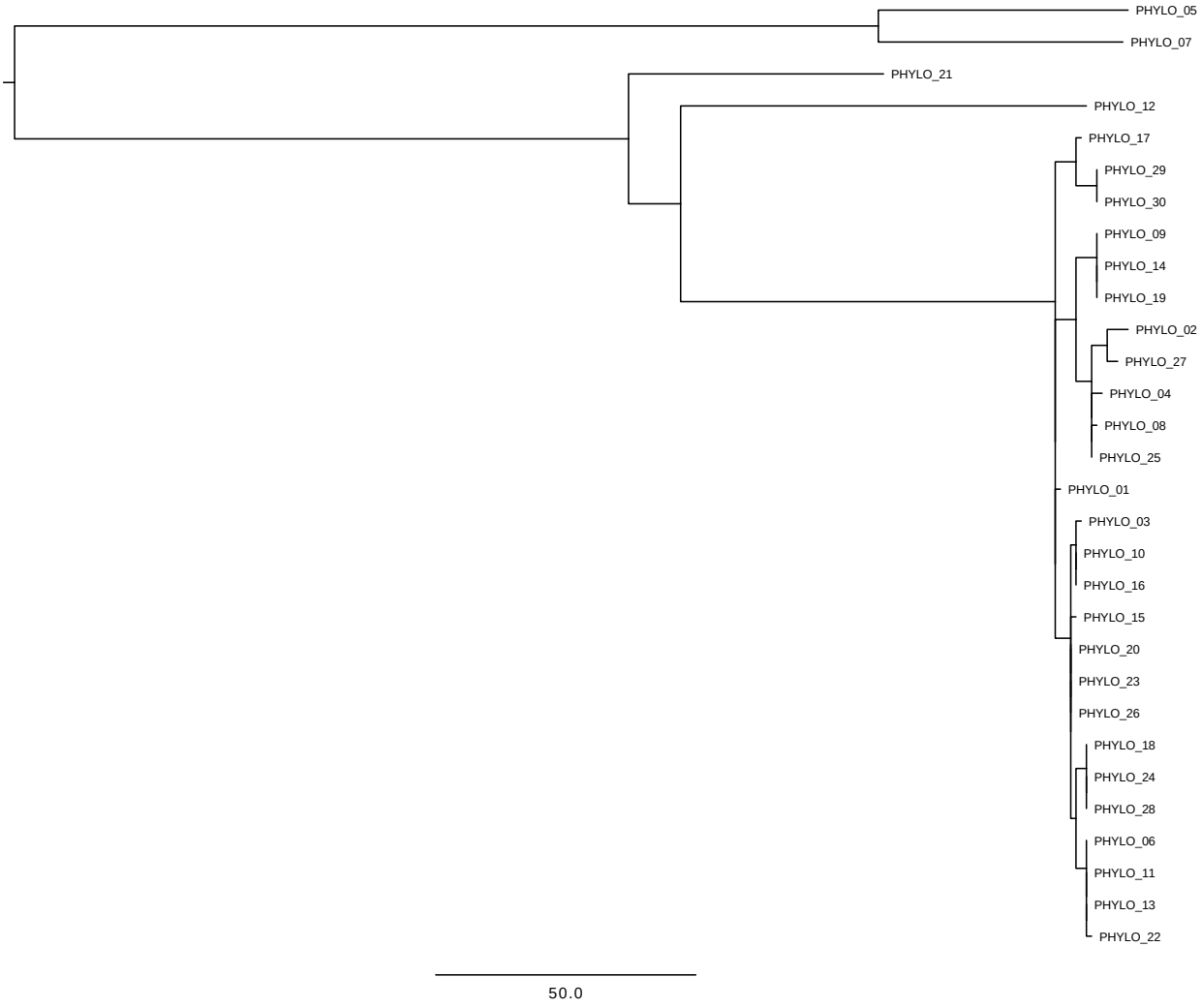
Scale represents the branch length stipulated into the newick file.

Figure I.7: Phylogeny Centre 3 obtained with CSI Phylogeny (CGE tools, command line version)



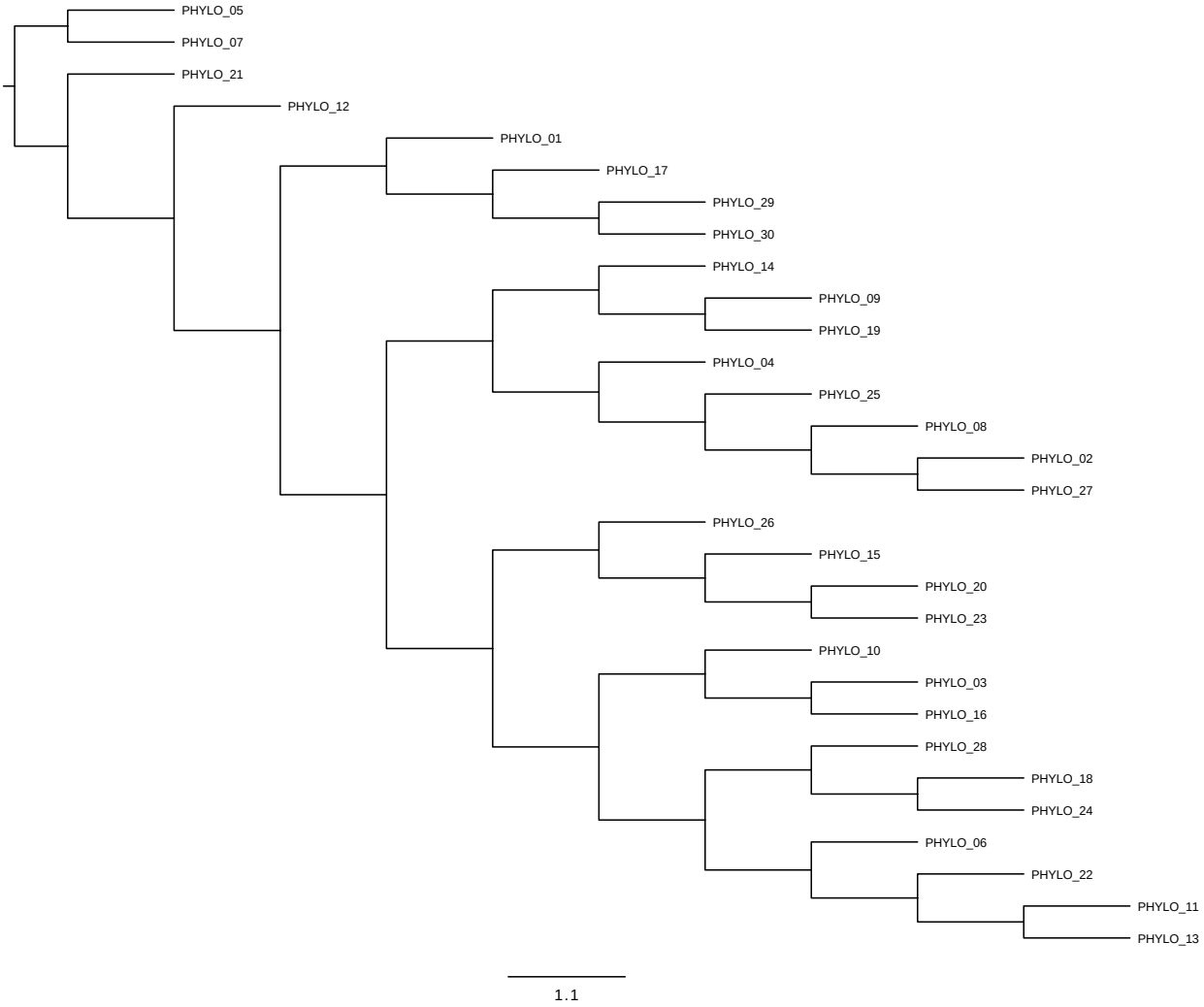
Scale represents the branch length stipulated into the newick file.

Figure I.8: Phylogeny Centre 4 obtained with CSI Phylogeny (CGE tools, online version)



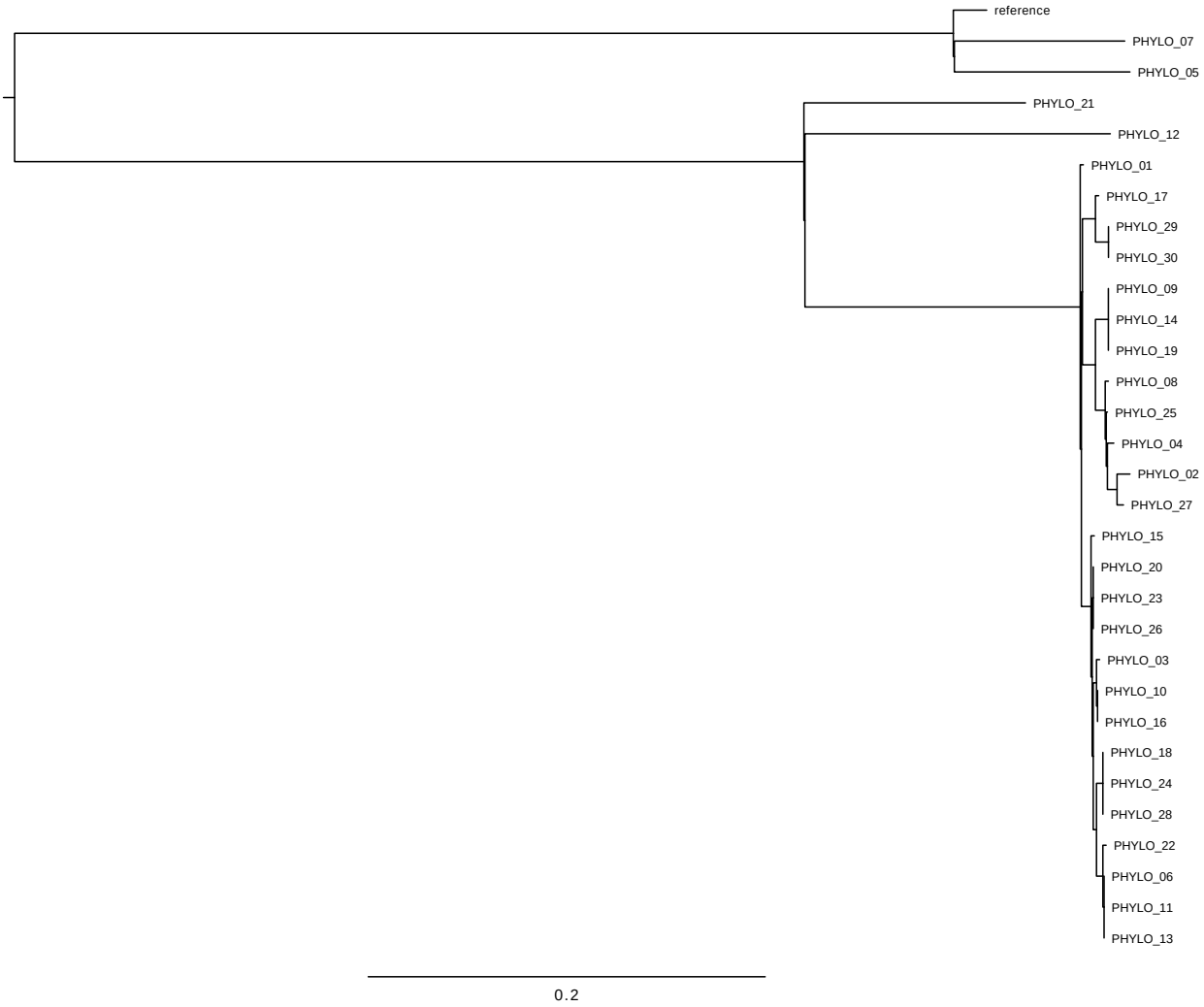
Scale represents the branch length stipulated into the newick file.

Figure I.9: Phylogeny Centre 5 obtained with CSI tools alignment (command line version) and a minimum evolutionary model for tree reconstruction



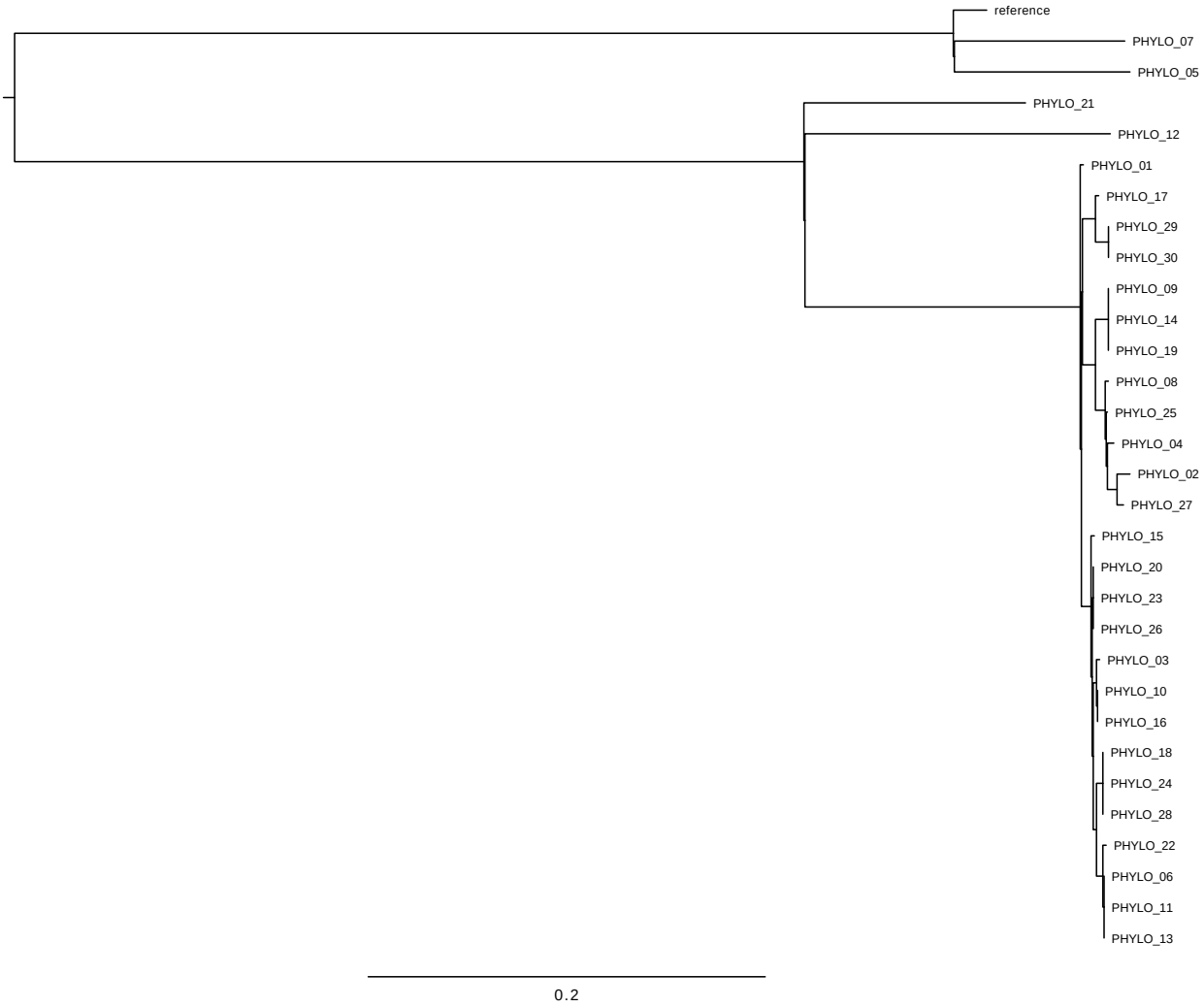
Due to some missing branch lengths in the newick format all the branches appear with the same length. Scale represents the branch length stipulated into the newick file.

Figure I.10: Phylogeny Centre 6 obtained with CSI Phylogeny (CGE tools, command line version) and a maximum parsimony tree reconstruction



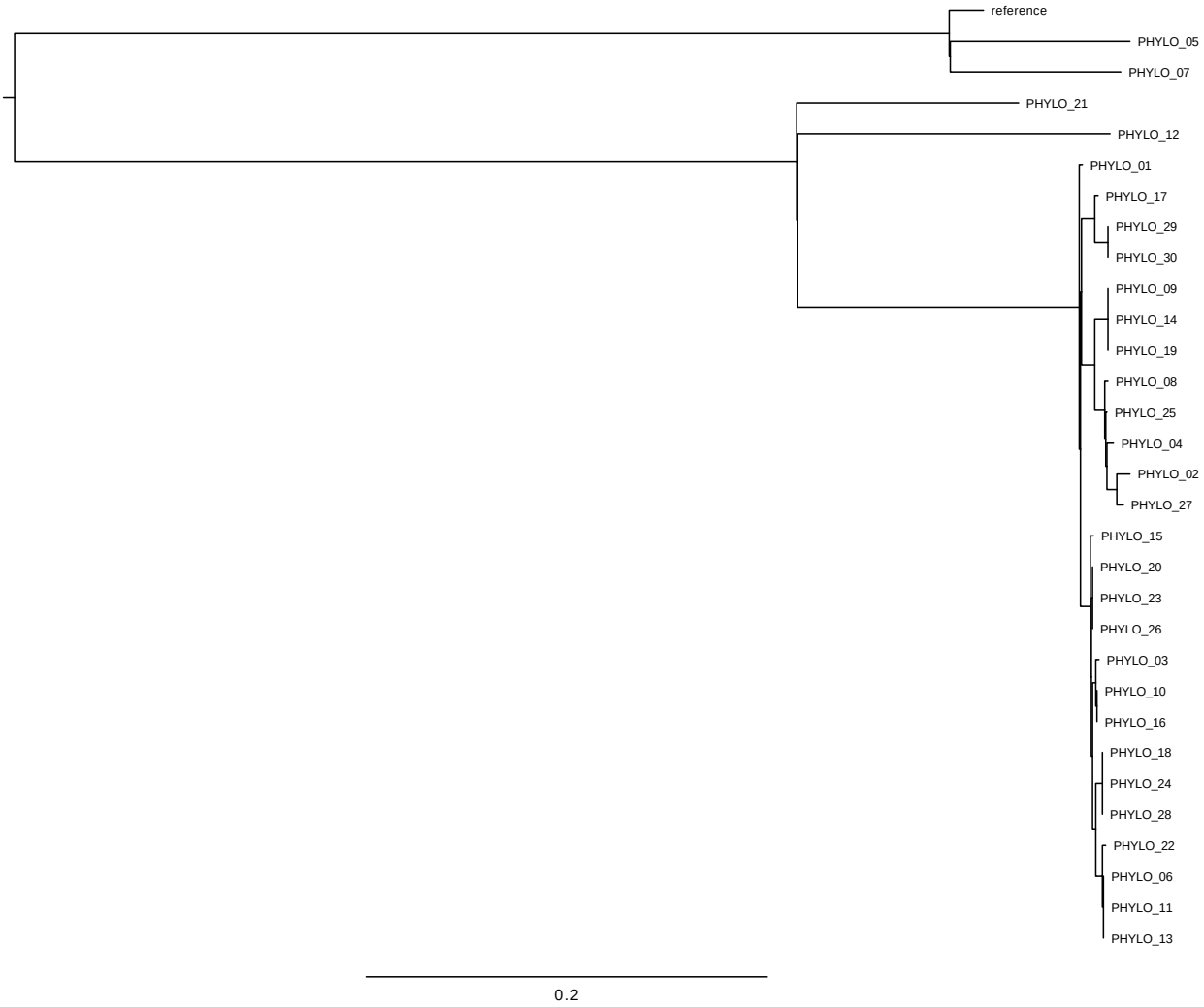
Scale represents the branch length stipulated into the newick file.

Figure I.11: Phylogeny Centre 7 obtained with CSI Phylogeny (CGE tools, online version)



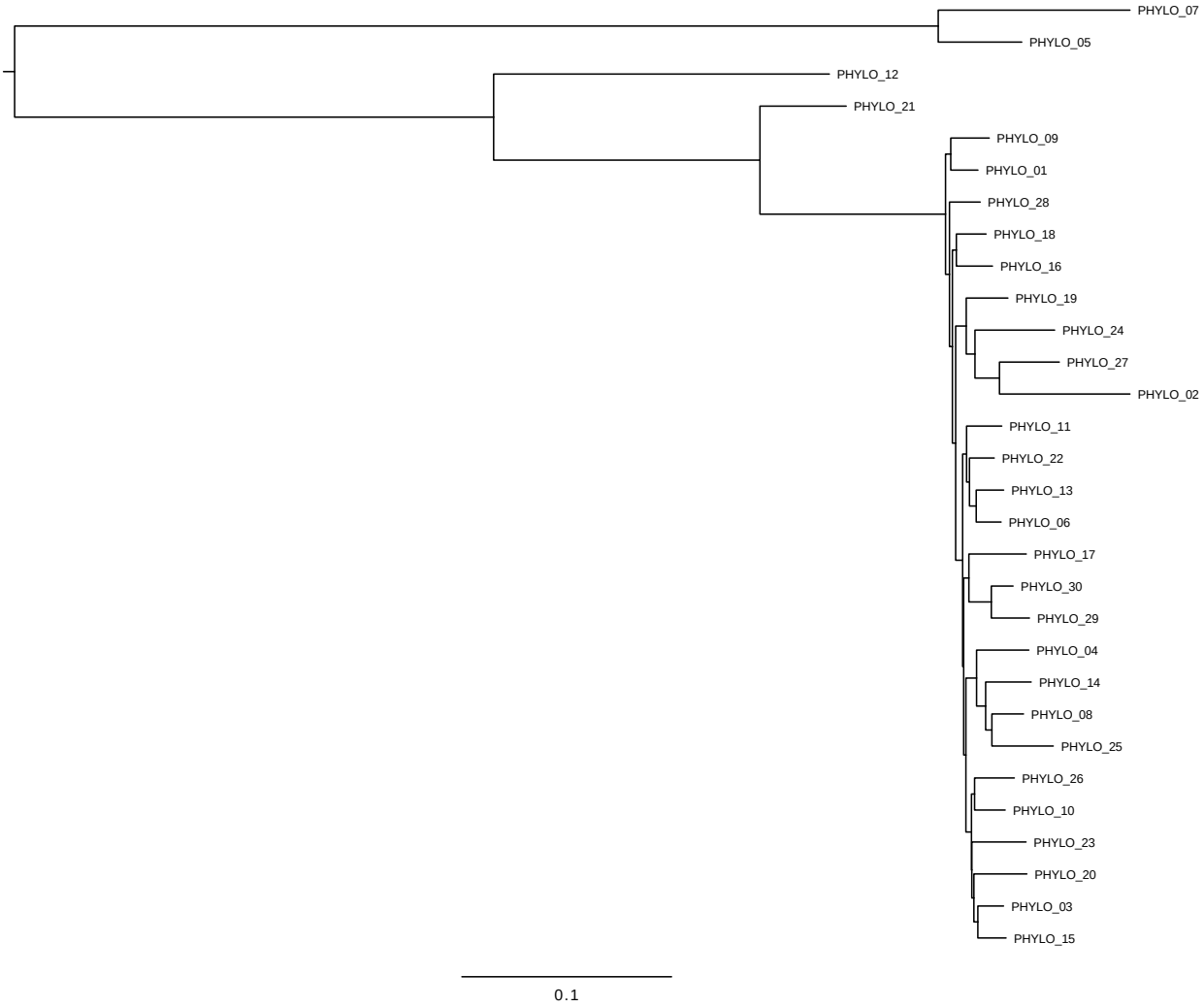
Scale represents the branch length stipulated into the newick file.

Figure I.12: Phylogeny Centre 8 obtained with CSI Phylogeny (CGE tools, online version) with heterozygous SNPs ignored



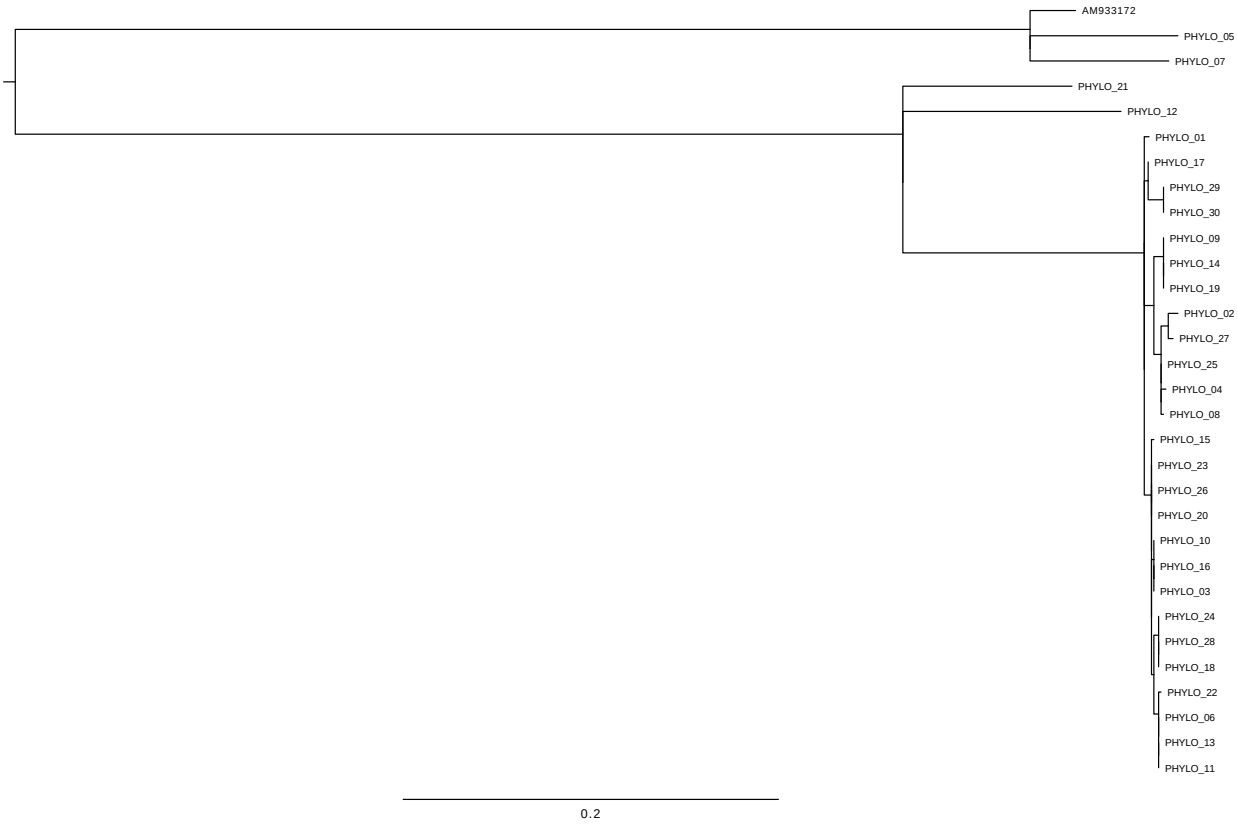
Scale represents the branch length stipulated into the newick file.

Figure I.13. Phylogeny Centre 9 obtained with CSI Phylogeny (CGE tools, online version)



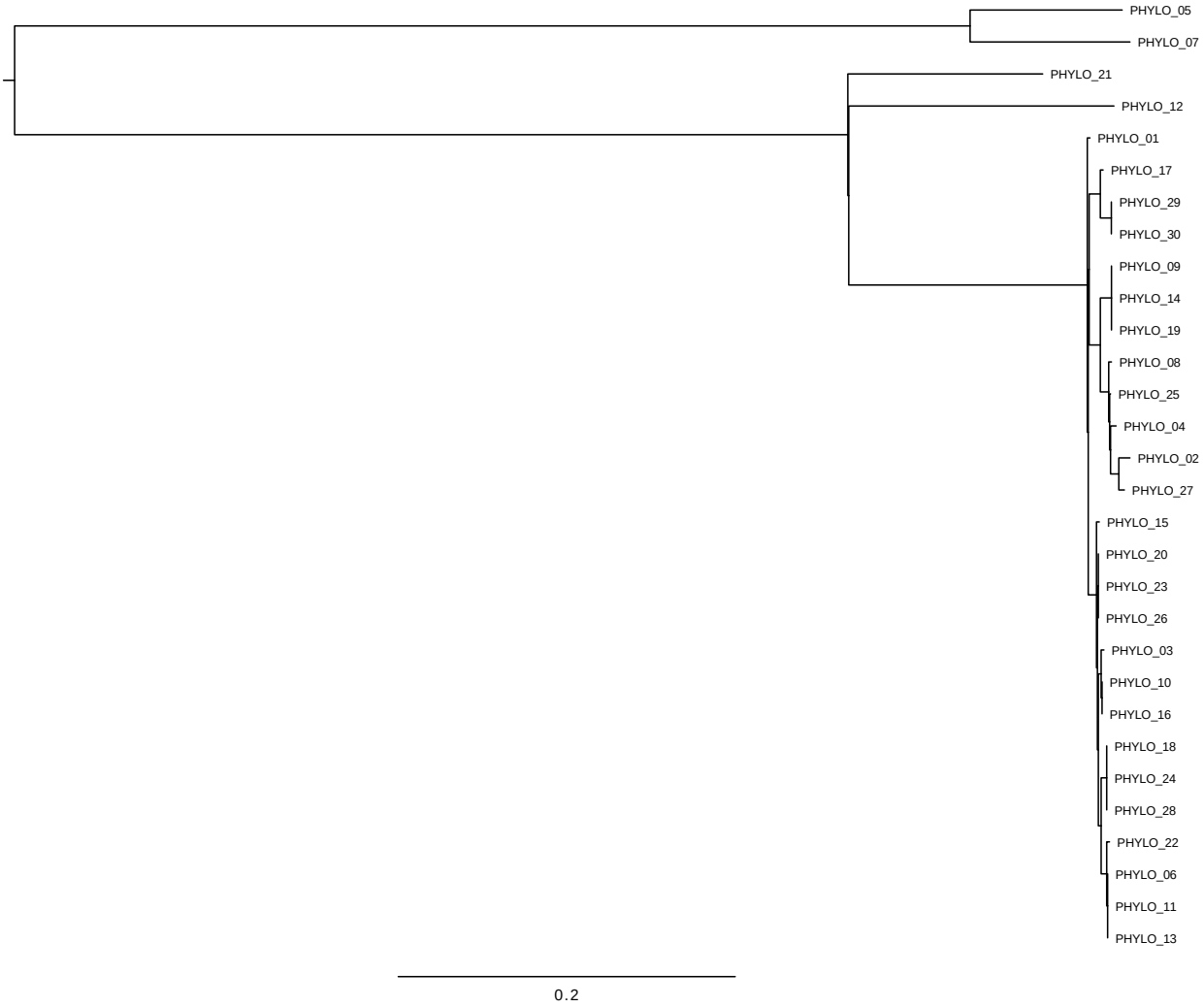
Scale represents the branch length stipulated into the newick file.

Figure I.14: Phylogeny Centre 10 obtained with BWA-Mem mapping, Freebayes, VariantAnnotation for detection/filter of SNPs and VCF-kit to generate the final alignment and the Neighbor Joining Tree



Scale represents the branch length stipulated into the newick file.

Figure I.15: Phylogeny Centre 11 obtained with PHEnix/SnapperDB and for variants detection and RAxML for tree building



Scale represents the branch length stipulated into the newick file.

Figure I.16: Partner Centre 12 phylogeny obtained with CSI Phylogeny (CGE tool, online version)

Appendix J – Benchmarking for *Campylobacter coli* phylogeny

Report number	#6
Responsible	Anaïs Painset (PHE) and Timothy Dallman (PHE)
Other partners/institutions involved	APHA (United Kingdom), BfR (Germany), DTU (Denmark), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland)
Benchmarking launched (date)	December 2017
Deliverable due (date)	January 2018

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools both to detect genomic variants and to build a phylogeny based on the variants detected for *Campylobacter coli* isolates. In this specific exercise, we ask participants to take into account the possibility of recombination between isolates. With the use of Whole Genome Sequencing, phylogenetic is used as a method to characterise microorganisms in outbreak investigations and for surveillance of isolates that are may be genetically related.

Participants

Participants in this benchmarking were institutions from the ENGAGE network.
Eleven sets of results were submitted from the following institutions:
APHA (United Kingdom), BfR (Germany), DTU (Denmark), IZSLT (Italy), IZSVe (3 phylogenies) (Italy), NIPH-NIH (Poland), NVRI (Poland), PHE (2 phylogenies) (United Kingdom).

Results from participating institutes are identified by codes (1-11, see below Table J.1) and each code is known only by the corresponding laboratory. The full list of laboratory codes is known only by the organizers (PHE). Table J.1 (below) describes the methods/tools used to produce each phylogeny. Final phylogenies will be referred as Centre XX later in the document where XX correspond to the row number of Table J.1.

Table J.1. List of tools/software used to produce each phylogeny submitted

Centre	SNP alignments tools (version) [parameters]	Tree building (version) [parameters]	Recombination detection (version) [parameters]
1	Snippy 3.0 [default]	Gubbins 2.1.0 [default]	Gubbins 2.1.0 [default] → post SNP detection
2	Snippy 3.2 [mapqual 60, basequal 20, mincov 10,minfrac 0.9] vcftools 0.1.15 [thin 100,recode]	FastTree 2.1.7 [-nt, Nucleotide distances: Jukes-Cantor, Joins: balanced, Support: SH-like 1000]	-
3	CSIPhylogeny 1.4 command line [default]	CSIPhylogeny 1.4 command line [default]	-
4	BWA Mem 0.7.12 [-p and default for other parameters] samtools 1.5 • view [-sb] • sort [default] • mpileup [-6 (Illumina +1.3), -C 50 (min quality 50), -v, -u] bcftools 1.5 • call [-O v, --ploidy 1, -v, -m] vcf_fa_extractor (https://github.com/moskalenko/vcf_fa_extractor) [default] clustalW built-in MEGA7 [default]	MEGA 7 [Statistical Method: Maximum likelihood, Model/Method: Jukes and Cantor, Rates among Sites: Uniforms, ML Heuristic method: NNI (Nearest-Neighbor-Interchange)	-
5	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] PHEnix 1.3 – bwa/GATKbuilt-in [sample_ploidy: 1, genotype_likelihoods_model: SNP, rf: BadCigar, out_mode: EMIT_ALL_SITES, nt: 1, ad_ratio: 0.9, min_depth: 15, qual_score: 30, mq_score: 30]	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-
6	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] Snippy 3.2	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-

Centre	SNP alignments tools (version) [parameters]	Tree building (version) [parameters]	Recombination detection (version) [parameters]
	[mincov 15, minqual 30, types snp]		
7	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] bwa-mem 0.7.12 [default] samtools 0.1.19-44428cd • mpileup [-I,-u] bcftools (part of samtools 0.1.19-44428cd) • view [-vcgI] vcfutils.pl varFilter [Q 25, d 30, w 10, W 15] Internal script for format conversion (vcf -> fasta)	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-
8	CGE CSI phylogeny online tools 1.4 [default]	CGE CSI phylogeny online tools 1.4 [default]	-
9	BWA mem 0.7.13 [default] Freebayes 1.1.0 [ploidy 1] VariantAnnotation 1.22.3 [remove monomorpic (AC > 0 & AC<8), set calls below 10x DP to no-call, remove variants with missing GT for >2 samples and no alts, remove variants with MQM < 50] VCF-kit 0.1.2 • pheno fasta	VCF-kit 0.1.2 • pheno tree nj	-
10	PHEnix 1.2 – bwa/GATK built-in [ad_ratio: 0.9, min_depth: 10, qual_score: 30, mq_score: 30] SnapperDB 0.2.4 [a: A80, r: Y, ng: gubbins gff]	RAxML 8.2.8 [-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT]	Gubbins 2.0.0 [c 16, u] recombination detected on WGS SNP alignment
11	PHEnix 1.2 – bwa/GATK built-in [ad_ratio: 0.9, min_depth: 10, qual_score: 30, mq_score: 30] SnapperDB 0.2.4 [a: A80, r: N, ng: gubbins gff]	RAxML 8.2.8 [-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT]	Gubbins 2.0.0 [c 16, u] recombination detected on WGS SNP alignment

Each row will be referred to as Centre XX.

Species/genomes included

Public Health England selected and provided genomes from *Campylobacter coli* and part of the same sequence type complex ST-828 complex. This ST-complex is very diverse as it is one of the most common ST-complex found amongst *Campylobacter* isolates. The genomes were selected because they were part of a suspected outbreak investigated by PHE. The outbreak occurred in 2008 in the North of England where a teacher and students from a primary school were having gastrointestinal

symptoms. A suspected tap water isolate was also collected. The isolates were sent for testing, all came back as *Campylobacter coli* and the same phage type PT44 as found. Retrospective WGS analysis shows that all the isolates were part of the same ST-complex: ST-828.

Nine genomes represented by sets of fastq (paired) were included the data set (Table J.6 with the list of selected genomes). All genomes were generated using an Illumina HiSeq and fastq provided to the partners were trimmed and assessed for quality before sharing. Fastq were trimmed using Trimmomatic 0.32 with the following options: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:8:true LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50. Quality of the sequencing was assessed by running FastQC 0.11.3. The trimmed and quality assessed reads were used for the analysis (see Table J.6 under 'Additional notes', and Supplementary Table 6 in Annex F).

In this outbreak, a recombination was suspected to occur; investigation confirmed the existing recombination that changes the topology and the branch length of the tree.

Two reference phylogenies were used; these were constructed by removing recombination region according to the gold standard methods. In the following report, gold standard methods used to generate the reference phylogenies consist of high quality SNPs filter, recombination detected, exclusion of the SNPs included into a recombination region and final phylogeny build with a maximum-likelihood method.

Tools used to build the reference phylogenies are PHEnix 1.2 for variants calling and filtering, Gubbins 2.0.0 for recombination detection, SnapperDB 0.2.4 to extract relevant SNPs and RAxML 8.2.8 to build the phylogeny. In this case reference phylogenies will be the phylogenies build following Centre 10 and Centre 11 tools/methods. One included the reference genome, the other did not. This choice was made to balance the bias related to the high diversity on the ST-complex.

Overall results

The results were compared using two main approaches:

1. Alignment and distance matrix comparison
2. Topology of the tree: global topology, Robinson-Fould symmetric difference and percentage of edge similarity (number of branches in one tree that are present in another)

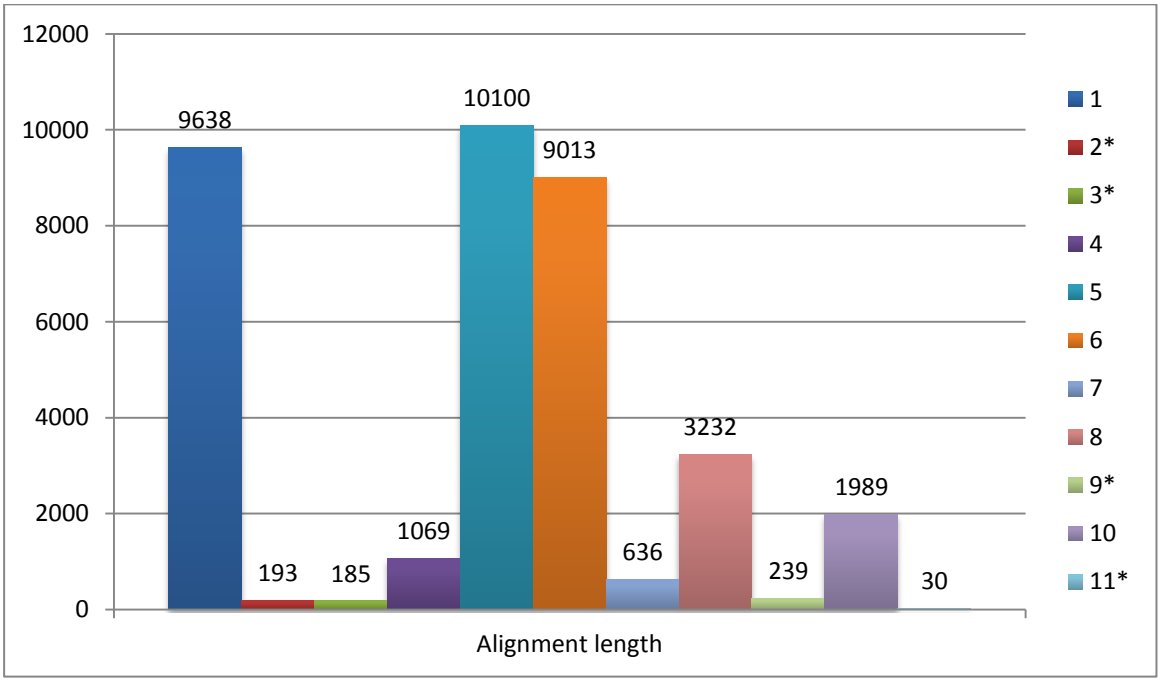
Each result was compared to their closest reference phylogeny i.e. with or without the reference genome.

Alignment and distance matrix

All the participants were required to provide a fasta alignment of the SNPs detected by the method they employed to generate the phylogeny. To ensure consistent comparison of the alignments, we generated the distance matrices from the alignment using an in-house python script. Distances from the reference phylogenies were calculate and the graphic generated using an R script.

Table J.2: Alignment and statistic metrics. Columns numbers correspond to Centre (ref. Table J.1)

	1	2	3	4	5	6	7	8	9	10	11
Min distance matrix	11	5	0	21	0	5	2	0	2	0	0
Max distance matrix	9367	157	177	212	9276	8804	6293	3121	230	1976	30
Reference	+	-	-	+	+	+	+	+	-	+	-



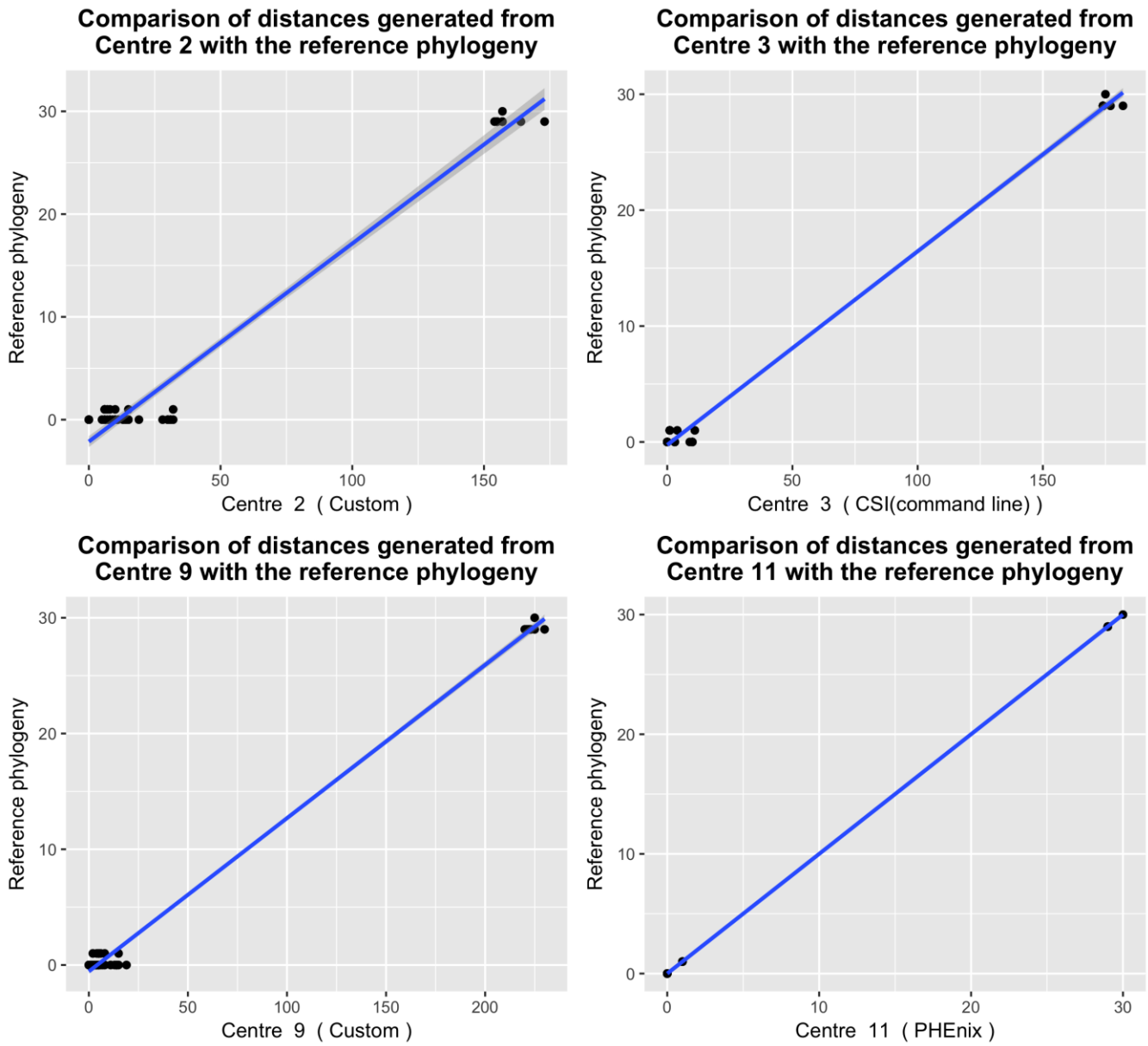
* indicate no reference in the alignment

Figure J.1: Size of the SNPs alignments for each result

For this benchmarking, we advised partners about potential recombination in the dataset. Results from Centre 10 and Centre 11 are taking into account recombination at the alignments step.

Results including the reference genome have significant longer alignments due to the number of SNPs present between samples from the dataset and the reference genome. *Campylobacter* is known for having huge diversity inside the same ST-complex. As we can see results that did not include the reference are including only SNPs detected between isolates of the dataset.

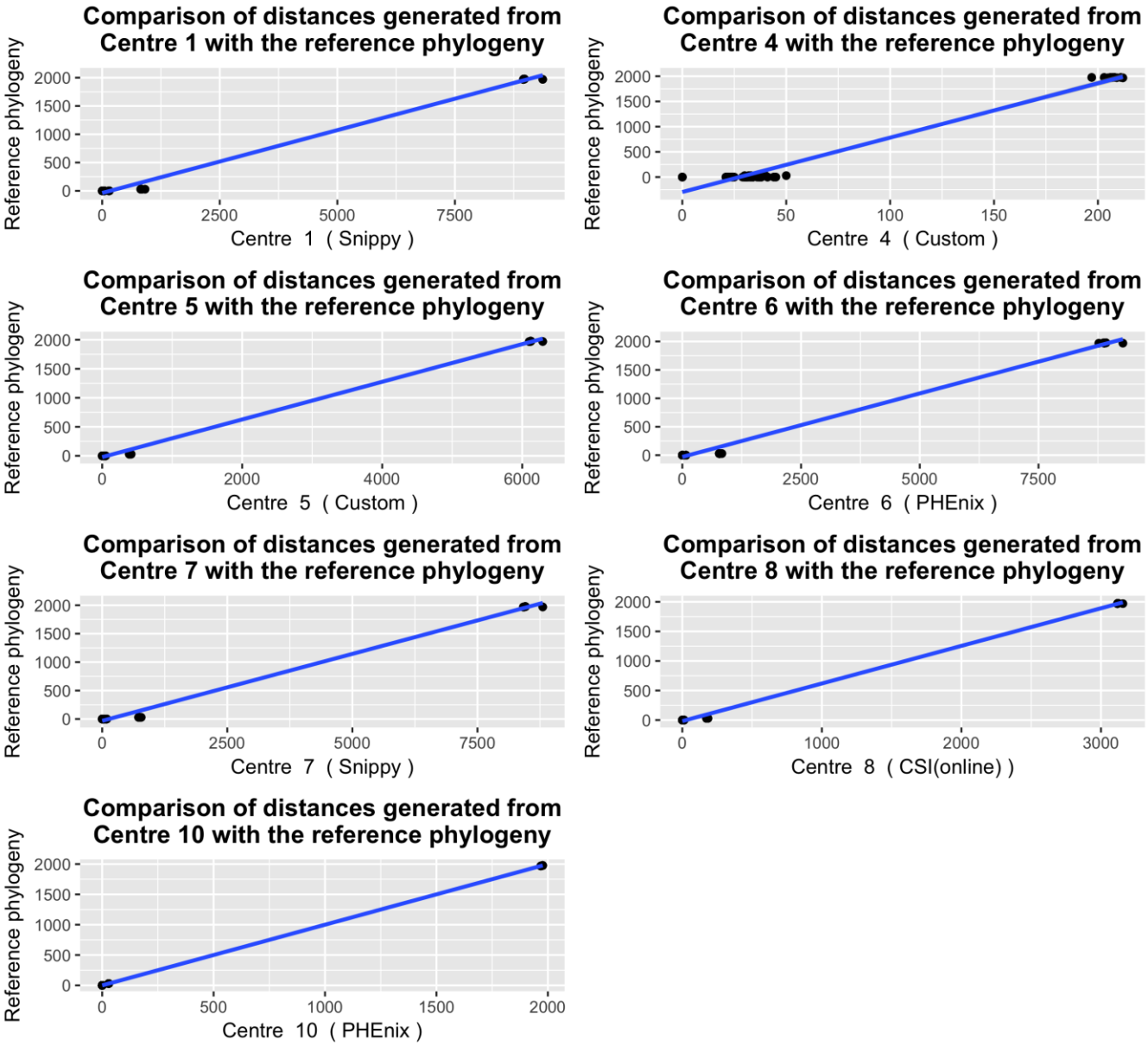
The selected isolates were part of a suspected outbreak and therefore the minimal SNP distance in the matrices should reflect the link between isolates. For eight out of eleven results the minimum SNPs distance is < 10 SNPs. This proves that all the methods are able to identify a strong link between some isolates of the dataset. The maximum distance found is highly related to the inclusion of the reference inside the alignment.



Centre numbers correspond to the list of benchmarking tools and participants for variants calling.

Figure J.2: Comparisons of distances generated from centre with gold standard without reference genome

The method used by Centre 2 and Centre 9 to generate the SNP alignment seems to show discrepancies for the closest isolates. This can be related to a recombination between closely related isolates.



Centre numbers correspond to the list of benchmarking tools and participants for variants calling.

Figure J.3: Comparisons of distances generated from centre with gold standard with reference genome

The methods used to generate the SNP alignment by the different partner show similar results except for Centre 4 where the comparisons of the distance matrix show discrepancies. There were also slight differences inside the closest isolates for all the methods used by the partners.

Topology of the tree

All the phylogenies are presented on the additional figures J.4 to J.16. They are labelled according to row number on the following table. The phylogenetic distance metrics were generated by using the ete toolktit (http://etetoolkit.org/) ete3 v.3.0.0 with his module compare and the additional phangorn R package v2.0.0.

Table J.3: Phylogenetic distance metrics

	ref*	E.size	nRF	RF	maxRF	%src-br+	%ref-br+	KF dist
1	+	10	0.86	12.00	14.00	0.62	0.62	25976.63
2	-	9	1.00	12.00	12.00	0.57	0.57	5.979280
3	-	9	0.78	7.00	9.00	0.82	0.64	5.037388
4	+	10	1.00	14.00	14.00	0.56	0.56	3.808749
5	+	10	1.00	14.00	14.00	0.56	0.56	1.428548
6	+	10	1.00	14.00	14.00	0.56	0.56	1.373172
7	+	10	1.00	14.00	14.00	0.56	0.56	1.174685
8	+	10	0.82	9.00	11.00	0.77	0.62	0.984009
9	-	9	1.00	12.00	12.00	0.57	0.57	9.104187
10	+	10	0.00	0.00	14.00	1.00	1.00	0
11	-	9	0.00	0.00	12.00	1.00	1.00	0

Row numbers correspond to Centre (ref. Table J.1)

* +/- indicated presence/absence of the reference in the final phylogeny

Additional notes: meaning of the metrics (ete-compare):

E.SIZE: effective size of the dataset used to calculate metrics

nRF: Normalized Robinson-Foulds distance (RF/maxRF)

RF: Robinson-Foulds symmetric distance

maxRF maximum Robinson-Foulds value for this comparison

%src_br (percent source branch): frequency of edges in target tree found in the reference (1.00 = 100% of branches are found)

%ref_br (percent reference branch): frequency of edges in the reference tree found in target (1.00 = 100% of branches are found)

KF.dist (Kuhner-Felsenstein distance): branch score distance (Kuhner & Felsenstein 1994) [compute with Phargorn]

The closer the normalized Robinson-Foulds (nRF) value is to 0, the better the match of the topology to the reference phylogeny. The results show that most of the trees are very different to the reference in terms of topology. Also, they seem to be consistently different from the reference phylogeny regardless of including the reference genome.

The closest phylogenies are from Centre 1, Centre 3 and Centre 8. Centre 3 and 8 used the CSI-phylogeny (CGE tools) and got better results. The phylogeny provided by Centre 1 is using a recombination detection software similar to the reference phylogeny explaining the similar results.

The KF distance measures the difference in term of branch length. As we can see most of the trees have somewhat a similar branch length. The Centre 1 branch length in the newick file has been derived from the recombination software used to build the phylogeny, it is not based on the SNP explaining why the KF distance (difference in term of branch length) is really high compared to the others methods shown.

Most of the partners have not used a specific tool or method to remove the recombination. Tables J.4 and J.5 show the matrix of KF distances comparing results between each other.

Table J.4: KF distance between all centre phylogenies – reference genome included

	1	4	5	6	7	8	10
1	0.00						
4	25980.44	0.00					
5	25978.06	2.38	0.00				
6	25978.00	2.44	0.06	0.00			
7	25977.81	2.64	0.25	0.20	0.00		
8	25975.65	4.79	2.41	2.35	2.16	0.00	
10	25976.63	3.81	1.43	1.37	1.17	0.98	0.00

Table J.5: KF distance between all centre phylogenies – reference genome not included

	2	3	9	11
2	0			
3	0.9487111	0.00		
9	3.2711648	4.16	0	
11	5.9792803	5.04	9.1041865	0

These matrices pointed out that some of the phylogenies are more similar in term of branch length between each other than they are with the reference phylogenies. For example phylogenies 5 and 6 seems highly similar.

Conclusion

The methods used to generate the SNP alignment by the different partners show similar results except for three (Centre 2, Centre 9 and Centre 4) where the comparisons of the distance matrix shows discrepancies between those isolates that are closely related. The distances matrices are informative to assess the relation between isolates and the phylogeny.

The overall topology of the trees compared to the gold standard reference is respected with all the methods able to pool together the isolates related to the outbreak and detach the suspected source from the main cluster. The main discrepancy linked to the presence of a recombination appears on the branch length and on the topology inside the closely related cluster.

The scores based on the topology demonstrate that most of the methods give different branch length and topology when a recombination occurs within the dataset, this could lead to biased distance matrices and an over-detection of SNPs. Removing the recombinant regions in this case shows strongest evidence of a cluster inside the dataset.

During this benchmarking we have identified that a key point in building a phylogeny where a recombination can occur is to link distance matrices and the phylogeny. The subtle topology of a closely related cluster is highly correlated to the presence of a recombination. We can also confirm with this benchmarking that if the organism is likely to contain recombination and discrepancies occurs between phylogeny and epidemiology information it is recommended to carry out a detection of recombination.

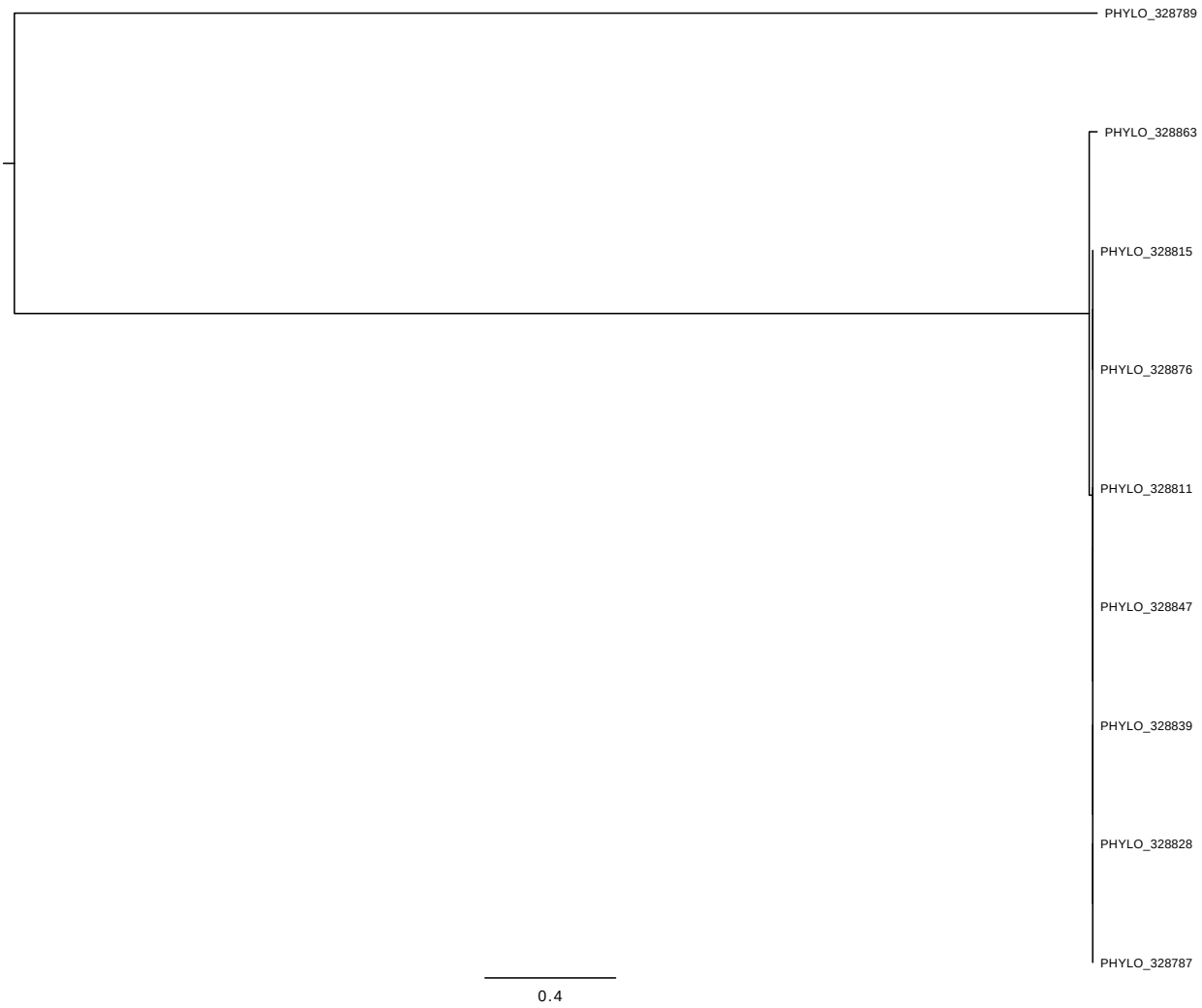
This benchmarking shows that partners have used "gold standard" methods both for SNP detection and tree building[12]. Filtering of SNPs has been properly carried out and most of the participants have used a maximum likelihood method to generate the phylogeny. It also demonstrated that despite knowing the good practise to derive a phylogeny from WGS, phylogeny need to be used with caution and can be only fully explained given support from other data, especially if in relation to outbreak investigations (i.e. for outbreak case definitions).

[12] Gold standard methods here referred to the filtering apply to detect SNPs and build the phylogeny with a maximum likelihood method.

Table J.6: Genomes selected for the benchmarking (further info in Supplementary Table 6 in Annex F)

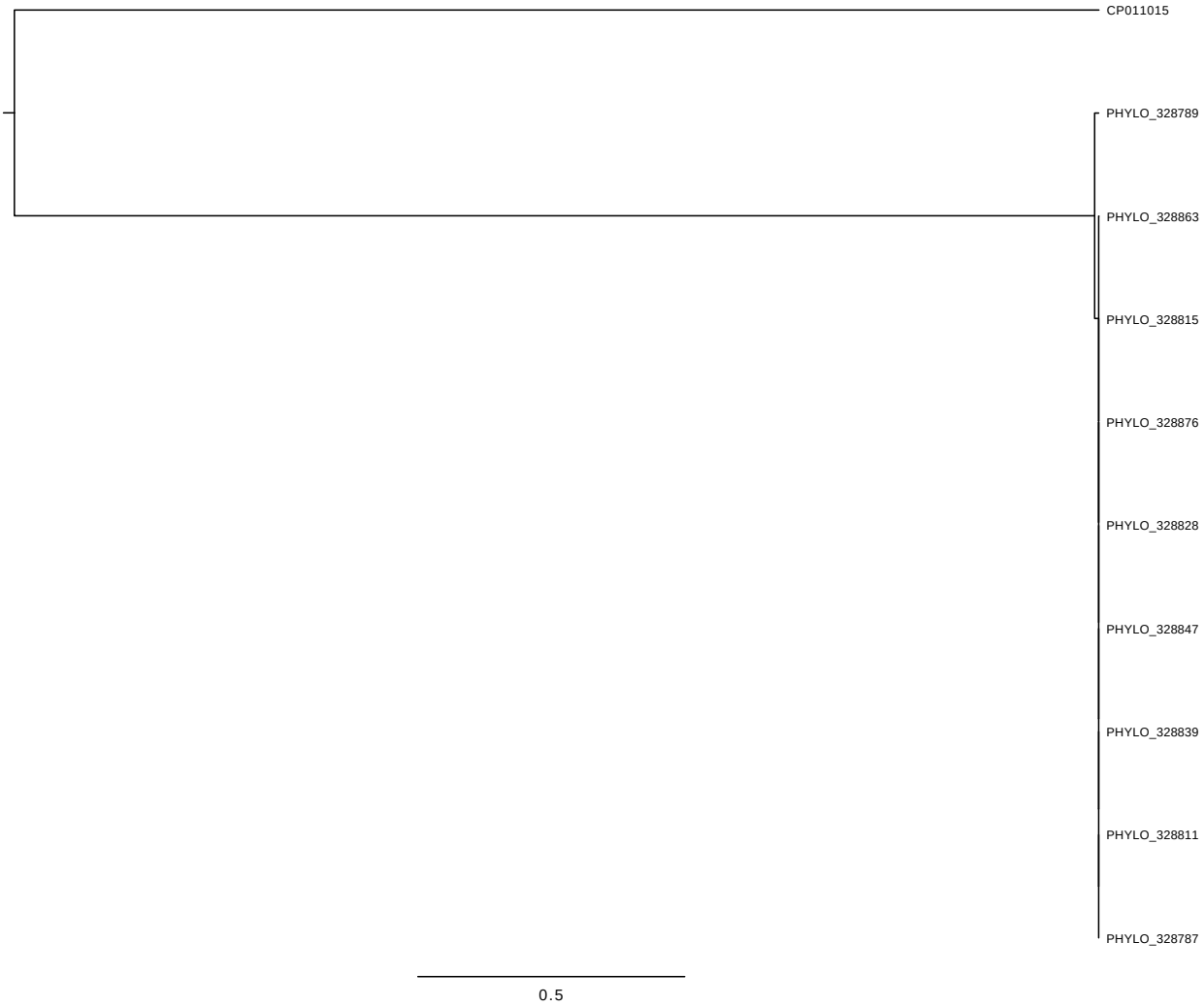
Sequence name
Reference: CP011015
PHYLO_CAMPY_328787
PHYLO_CAMPY_328789
PHYLO_CAMPY_328811
PHYLO_CAMPY_328815
PHYLO_CAMPY_328828
PHYLO_CAMPY_328839
PHYLO_CAMPY_328847
PHYLO_CAMPY_328863
PHYLO_CAMPY_328876

Additional figures



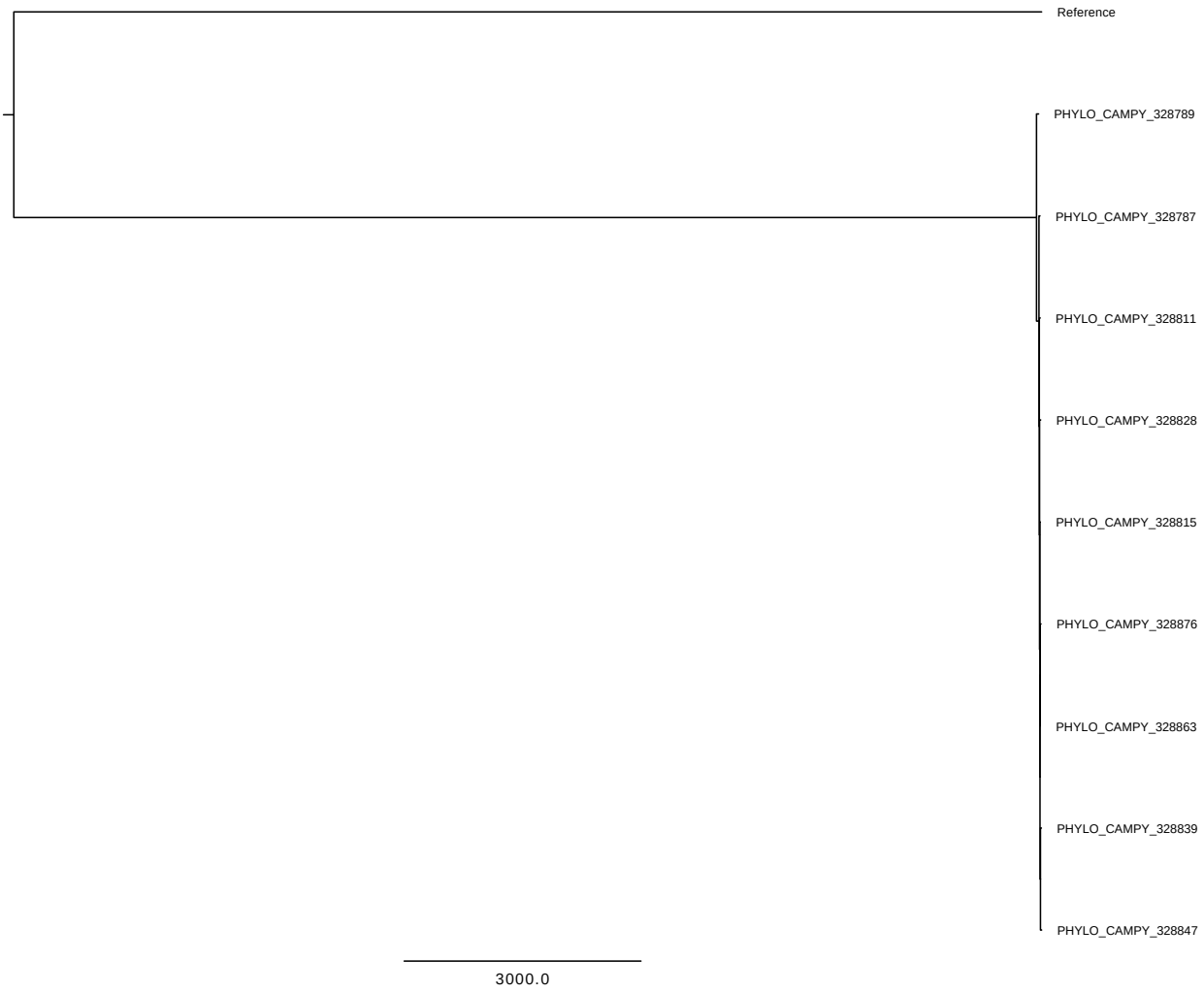
Scale represents the branch length stipulated into the newick file.

Figure J.4: Reference phylogeny without reference genome and recombination removed



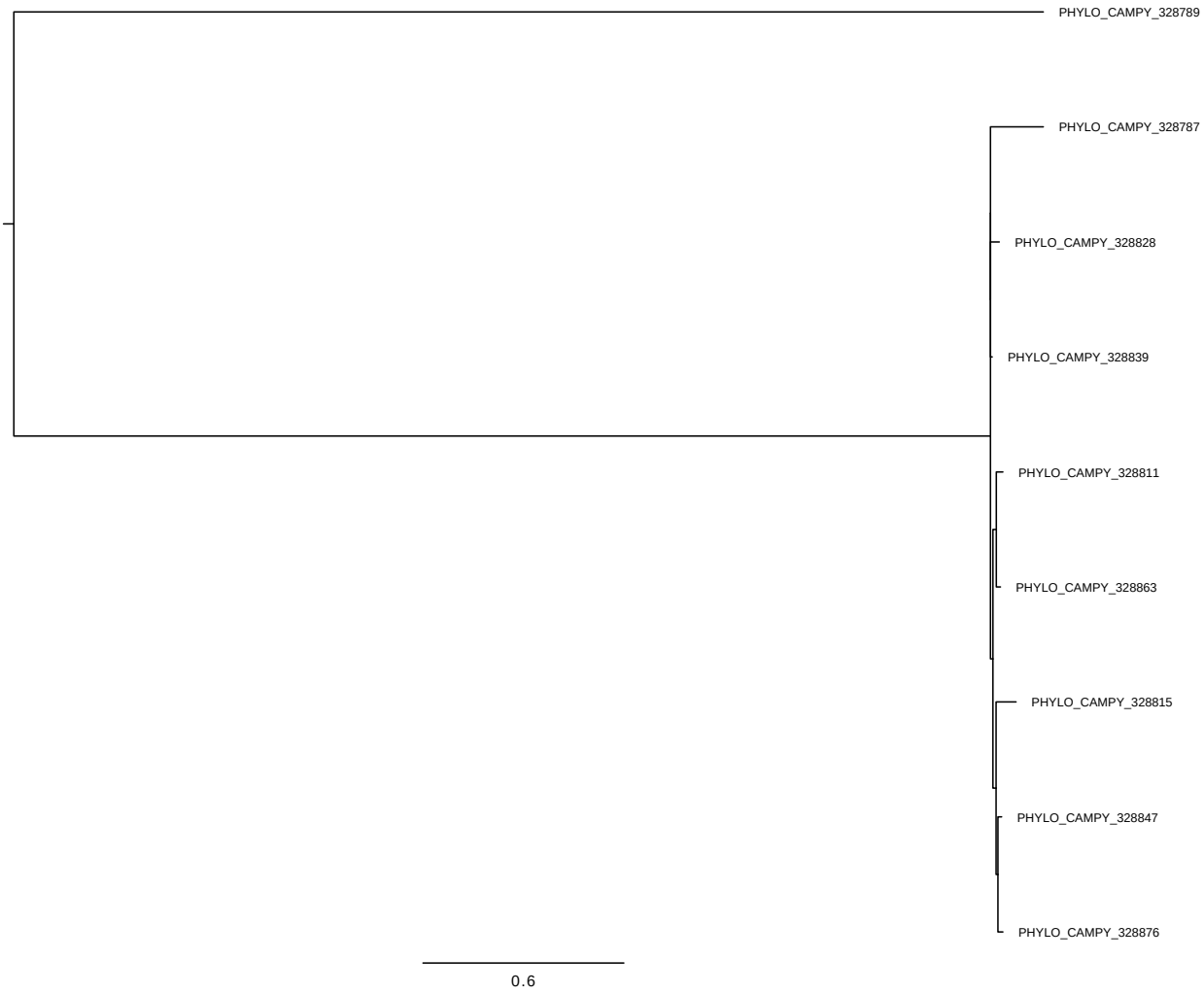
Scale represents the branch length stipulated into the newick file.

Figure J.5: Reference phylogeny with reference genome and recombination removed



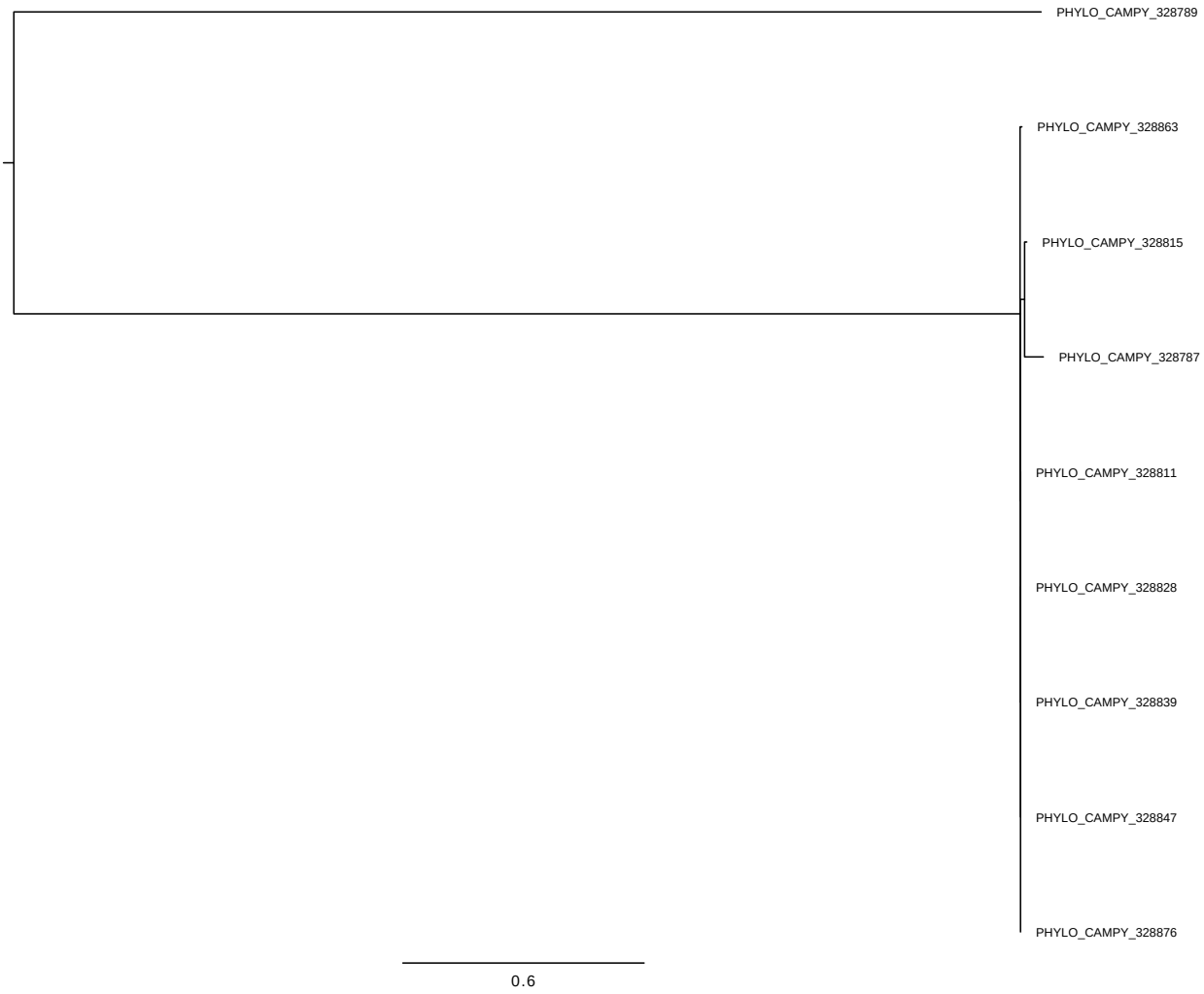
Scale represents the branch length stipulated into the newick file.

Figure J.6: Phylogeny results Centre 1



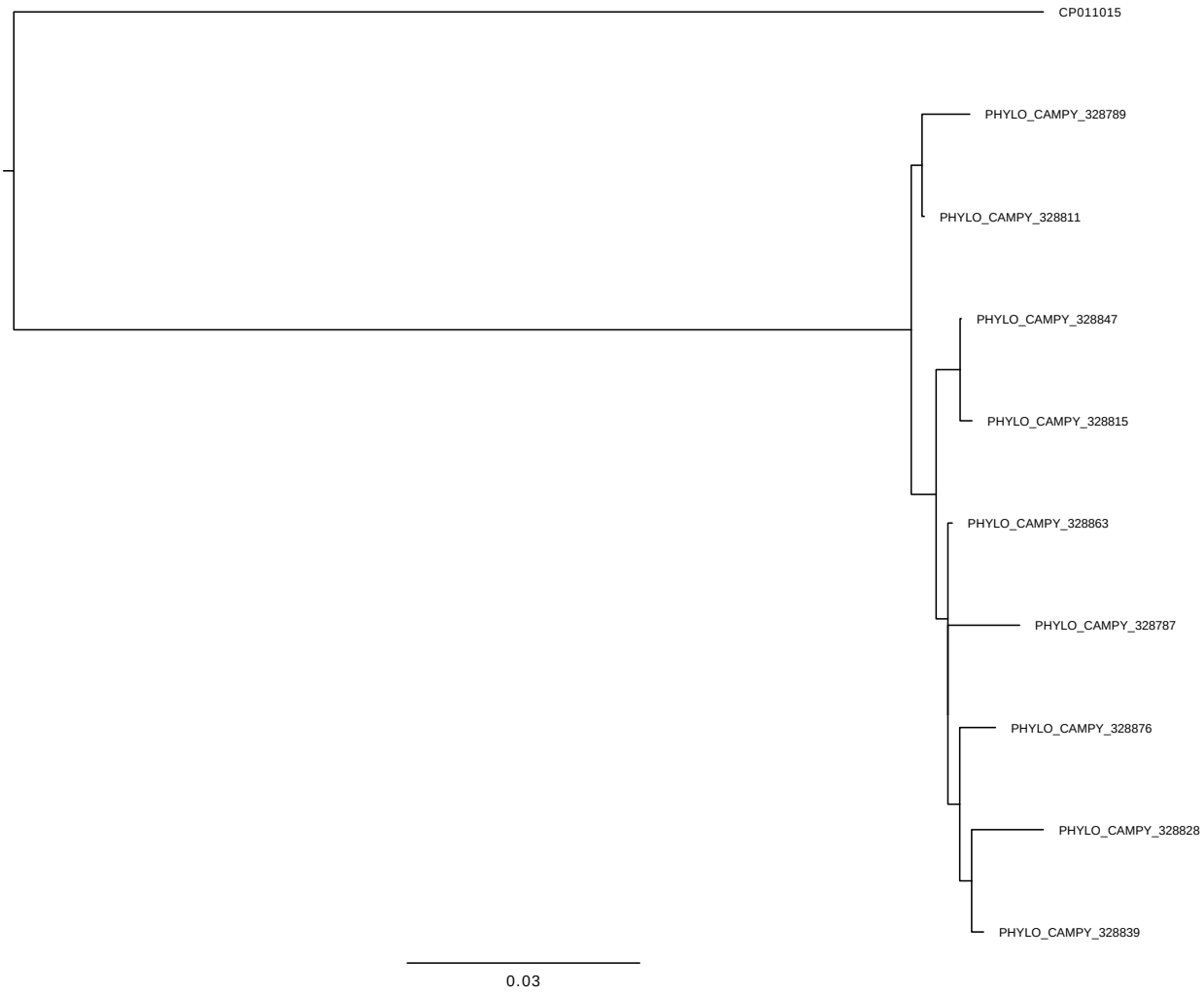
Scale represents the branch length stipulated into the newick file.

Figure J.7: Phylogeny results Centre 2



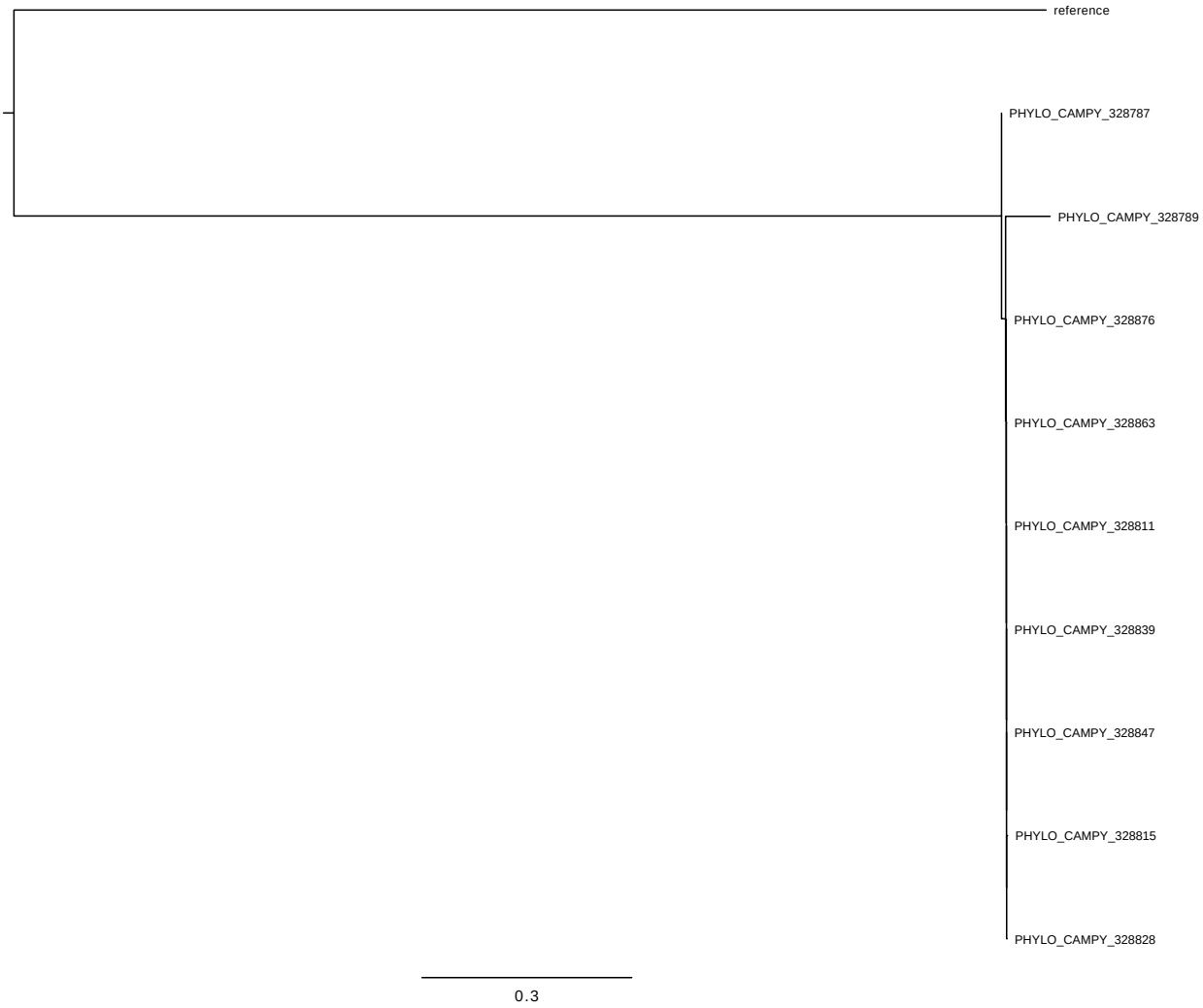
Scale represents the branch length stipulated into the newick file.

Figure J.8: Phylogeny results Centre 3



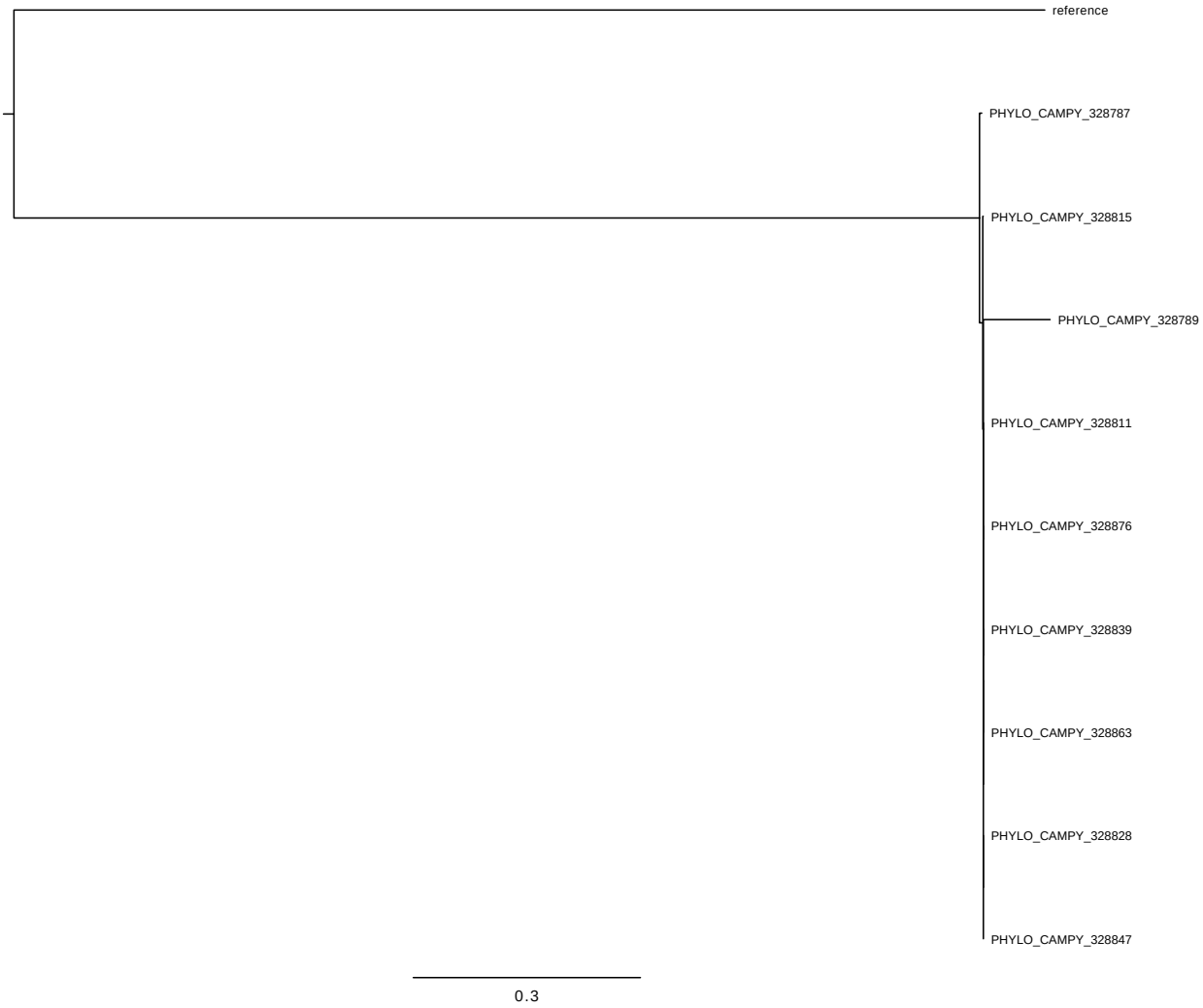
Scale represents the branch length stipulated into the newick file.

Figure J.9: Phylogeny results Centre 4



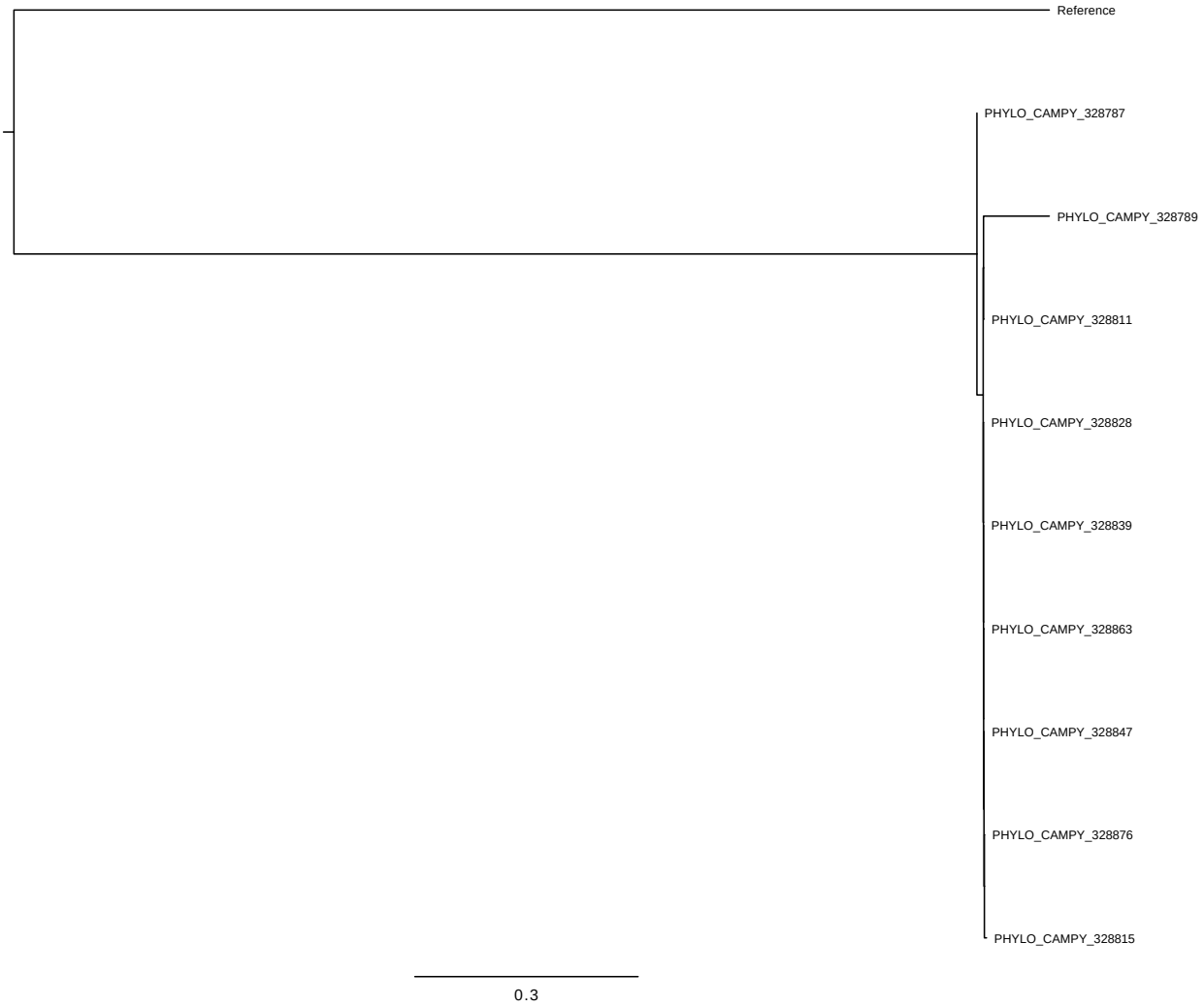
Scale represents the branch length stipulated into the newick file.

Figure J.10: Phylogeny results Centre 5



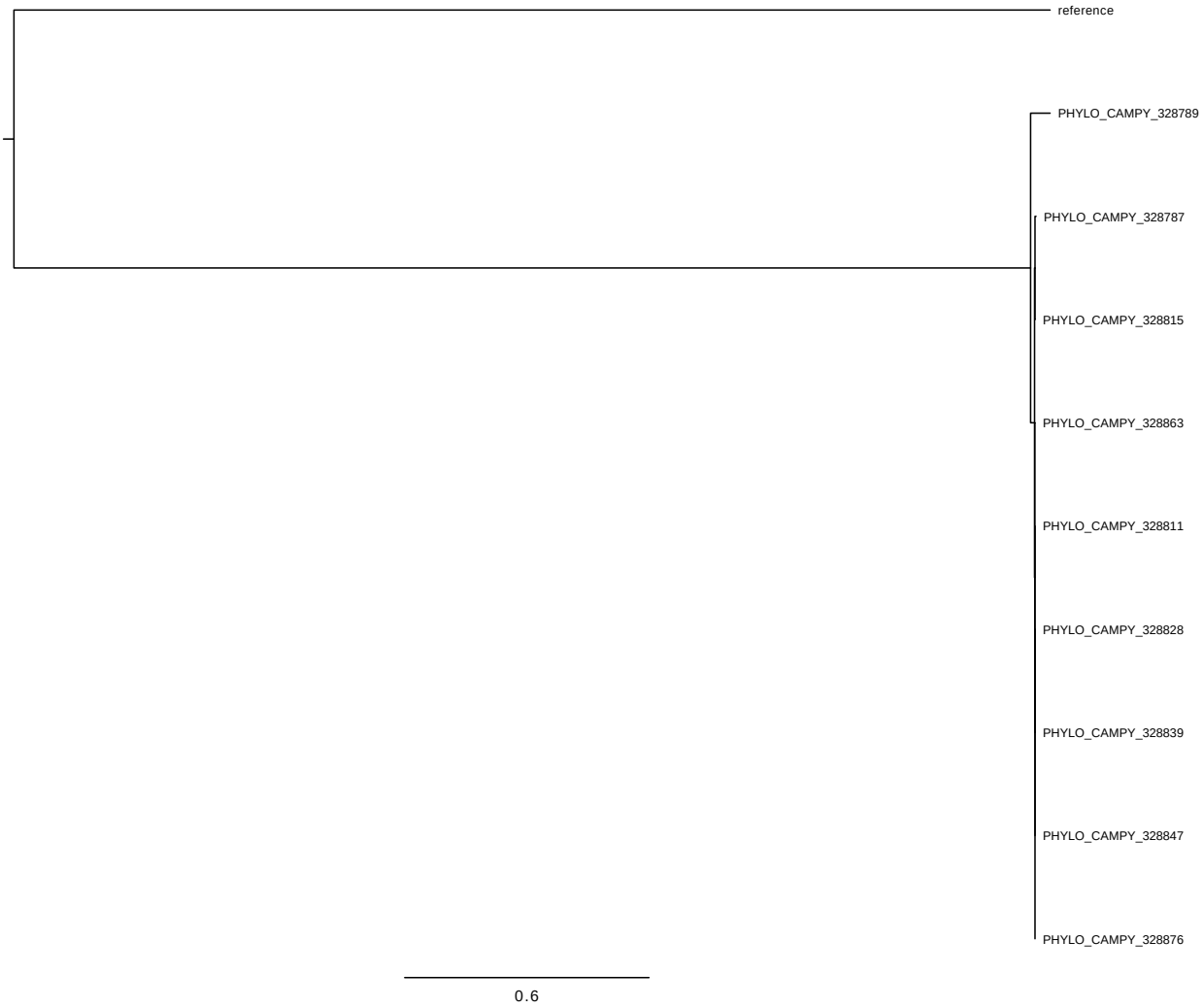
Scale represents the branch length stipulated into the newick file.

Figure J.11: Phylogeny results Centre 6



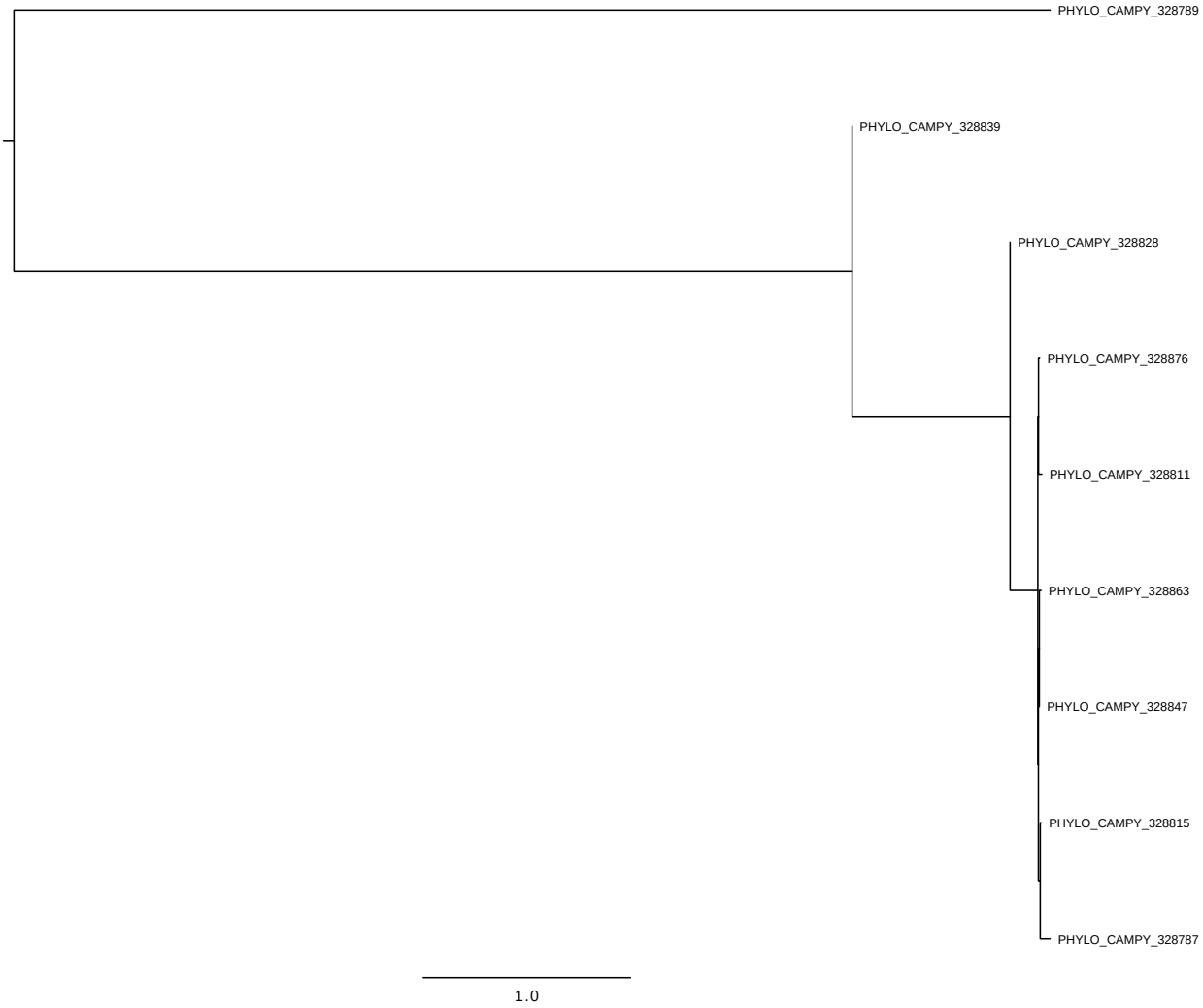
Scale represents the branch length stipulated into the newick file.

Figure J.12: Phylogeny results Centre 7



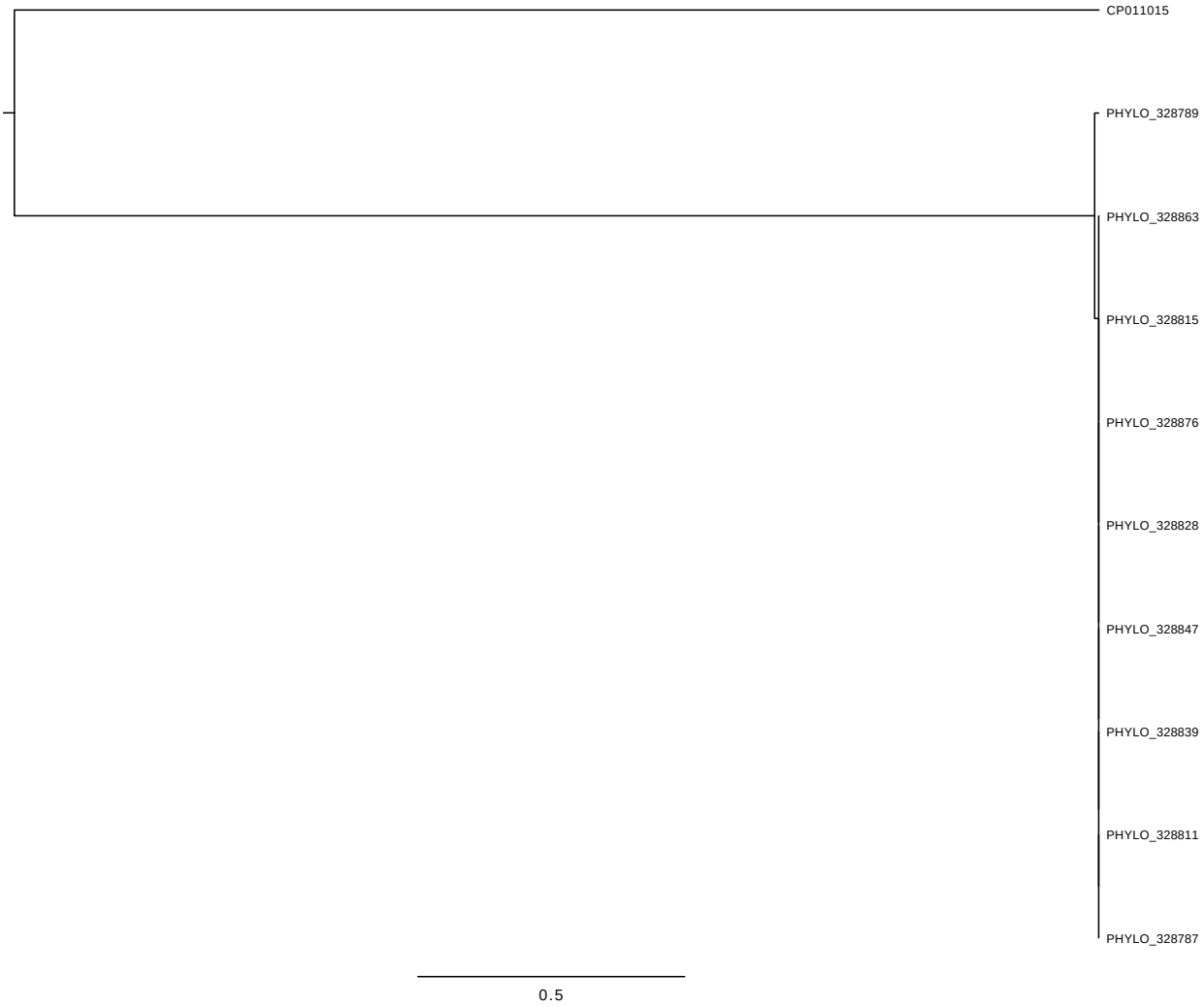
Scale represents the branch length stipulated into the newick file.

Figure J.13: Phylogeny results Centre 8



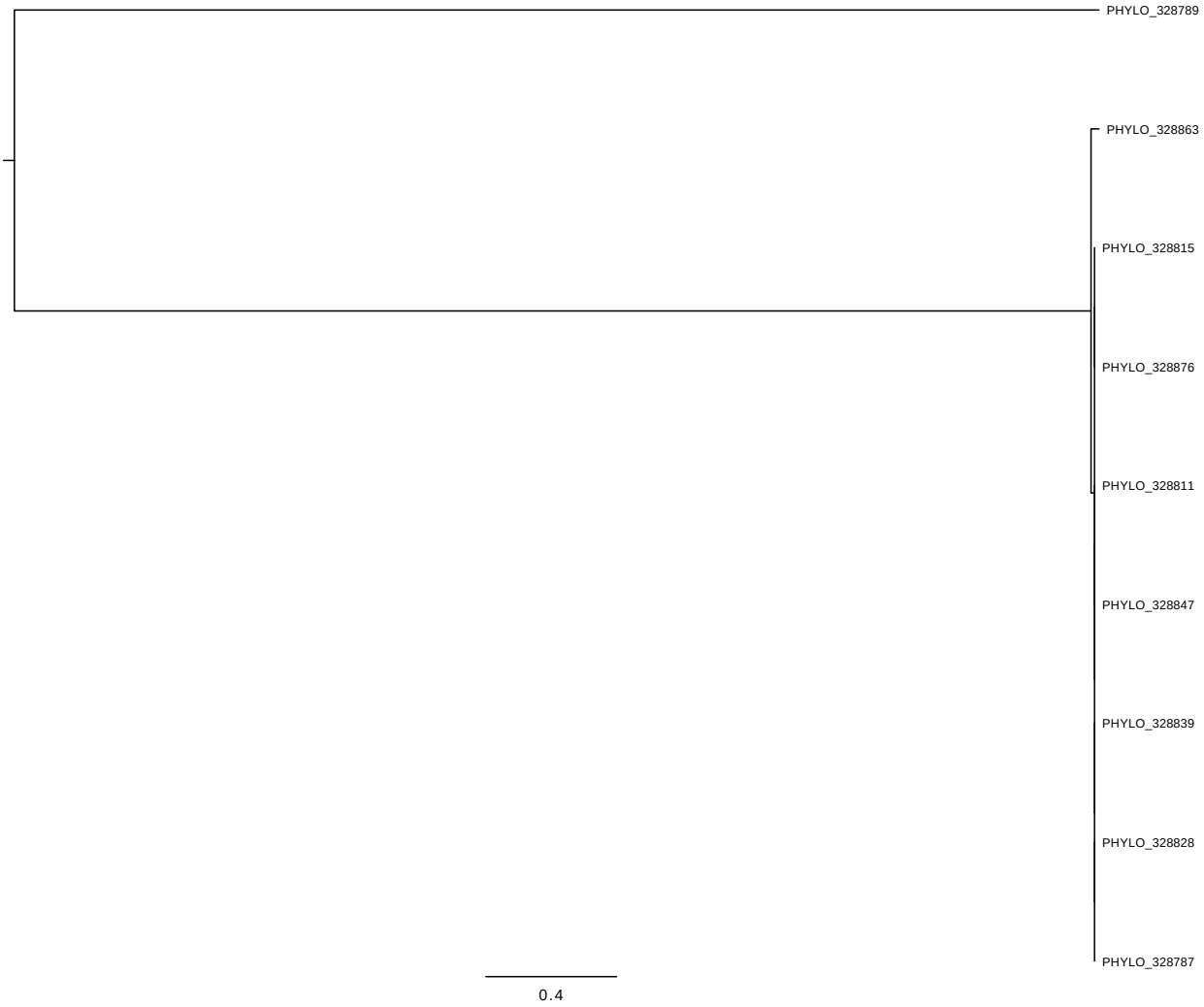
Scale represents the branch length stipulated into the newick file.

Figure J.14: Phylogeny results Centre 9



Scale represents the branch length stipulated into the newick file.

Figure J.15: Phylogeny results Centre 10



Scale represents the branch length stipulated into the newick file.

Figure J.16: Phylogeny results Centre 11

Appendix K – Proof of concept projects

Appendix K.1

Project title	*mcr-1*-harbouring plasmids in *Escherichia coli* in Denmark and their phylogenetic relationship with *mcr-1* plasmids from other geographical regions
Consortium partner	DTU
Scientific lead	Valeria Bortolaia, Jette Sejer Kjeldgaard, Ana Rita Rebelo, Pimlapas Leekitcharoenphon, Rene Hendriksen
Other people involved	Helle Bisgaard Korsgaard, Hanne Mordhorst, Frank Møller Aarestrup

Purpose

To assess possible pathways of dissemination of *mcr-1* in *E. coli*

Background

Plasmid-mediated colistin resistance is an emergent public health threat since colistin is used as last resort drug to treat multidrug-resistant infections by *E. coli* in humans.

Although detection of *mcr-1* in Enterobacteriaceae has been extensively reported worldwide, few studies performed in depth characterization of the plasmids harbouring *mcr-1*.

Objectives

The objective of this study was to characterize *mcr-1*-harbouring plasmids in *E. coli* in Denmark.

Project description

Currently, the DANMAP collection at DTU holds 13667 *E. coli* isolates from poultry (n=6808), pig (n=4302) and calf (n=2557) collected between 2001 and 2016. Of those, 5447 strains are resistant to colistin with an MIC ≥ 4 mg/L while the remaining 8220 are susceptible isolates with MIC ≤ 2 mg/L. Our test collection consisted on two subsets extracted from this assortment and was selected based on MIC values and/or year of collection. A total of 115 isolates from food or animal origin were screened from which 50 were resistant to colistin with MIC ≥ 4 mg/L and 65 were susceptible with a MIC of 2 mg/L (Figure 1):

- The subset of resistant strains corresponded to all isolates with MIC ≥ 4 mg/L collected for DANMAP between 2008 and 2013, composed by samples of poultry (n=45), swine (n=4) and calf (n=1). Six samples were from Denmark and 44 were of international origin;

- The subset of susceptible isolates corresponded to all *E. coli* strains with MIC of 2 mg/L collected since 2001, obtained from poultry (n=37), pig (n=14) or calf (n=14). 39 strains were of national origin and the remaining 26 were from imported products.

Methods and Results

PCR analysis revealed a total of 51 *mcr-1*-positive isolates. Of these, 41 corresponded to colistin resistant strains (MIC ≥ 4 mg/L) and 10 were colistin susceptible strains (MIC = 2 mg/L) (Figure 1). All isolates harboring *mcr-1* were samples from poultry, from which 50 were of international origin and one was from Denmark.

Whole genome sequencing was performed on the 51 *mcr-1*-positive isolates and on the remaining 9 colistin resistant isolates (41 strains were sequenced under ENGAGE project), and the analysis was performed as follows:

- *mcr-1*-positive isolates were screened for co-occurrence of *mcr-2, mcr-3, mcr-4* and *mcr-5*. All isolates were negative for other *mcr* genes.
- Colistin resistant but *mcr*-negative isolates were investigated *in silico* for other mechanisms of colistin resistance. One national swine isolate with a MIC of 8 mg/L presented a point mutation in *pmrA/pmrB* two component system (*pmrB* V161G). The lasting eight resistant isolates, with MIC between 4 mg/L and 8 mg/L, did not present any known mechanisms of resistance (Figure K.1).
- Phylogeny of the *E. coli* harbouring the *mcr-1* plasmids was inferred. Of the 60 isolates, 43 were successfully characterized with both ST and serotype, nine were characterized with ST but not serotype, three were characterized with serotype but not ST and five were undetermined for both features. All strains were analyzed by cgMLST. The 60 strains belonged to 36 different STs, with eight remaining undetermined. The most prevalent ST was ST-10 (n=6). Using SerotypeFinder we were able to characterize 46 isolates (14 were undetermined), which belonged to 40 different serotypes. cgMLST revealed 53 different core-genome sequence types and the construction of a phylogenetic tree based on these allowed for clustering of strains of the same STs and serotypes. Three strains were removed from the tree due to the low extent of allele attribution (353, 1666 and 331 undetermined alleles out of 2513, respectively). All three corresponded to colistin resistant strains without known mechanisms of resistance.
- All isolates presented varied profiles of antimicrobial resistance, virulence factors and plasmid content as determined by CGE batch upload pipeline.
- The contigs harbouring *mcr-1* were used to determine plasmid incompatibility type. 19 isolates have the gene integrated in IncX4 plasmids, one isolate has the gene integrated in the chromosome, and the remaining isolates possess the gene integrated either in type IncH, IncP or IncI (further analysis is necessary to determine incompatibly types).

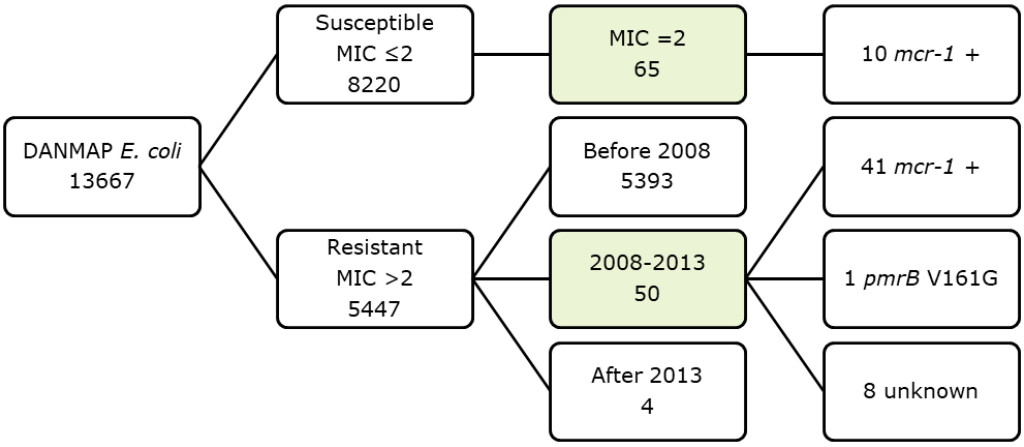


Figure K.1: Isolate selection process and colistin resistance mechanisms. Highlighted the subsets of strains included in the studies

Future work includes:

- Determination of *mcr*-harbouring plasmid incompatibility type for the remaining isolates;
- Annotation of *mcr-1*-positive contigs to examine the *mcr-1* genetic context;
- Single Nucleotide Variation-based phylogeny of the annotated *mcr-1*-positive contigs for each plasmid type detected;

Comparison with the genetic context of *mcr-1* harboured in plasmids of the same types described in other geographical locations.

Additional notes

Manuscript draft planned for June 2018

Appendix K.2

Project title	Phylogeny of *Salmonella* Paratyphi B variant Java (ST-28) harbouring *mcr-1*
Consortium partner	BfR
Scientific lead	Burkhard Malorny, Maria Borowiak
Other people involved	Rene Hendriksen, Pimlapas Leekitcharoenphon

Purpose

The study provides new insights into the evolution and spread of plasmid-mediated colistin resistance in *S. enterica* that can be useful for reduction strategies on resistance development in animals and humans.

Background

Liu and colleagues (2015) have recently described a mobilizable colistin resistance gene, *mcr-1,* located on a plasmid. The authors observed *mcr-1* carriage in China in *Escherichia coli* isolates in 15% of 523 samples of raw meat, in 21% of 804 samples of animals and in 1% of 1322 samples from patients with infection during 2011 to 2014. Meanwhile, a number of publications reported also the occurrence of *mcr-1* in European isolates from food, animals and humans, especially in *E. coli* and *S. enterica.*

Objectives

Objectives were to study the incidence and microevolution of *mcr-1* in *d*-tartrate fermenting (*d*Ta+) *S. enterica* subsp. *enterica* serovar Paratyphi B (variant Java) belonging to the sequence type (ST) 28 in terms of phylogeny of isolates carrying *mcr-1* on plasmids and the horizontal spread of *mcr-1* throughout the food chain.

Project description

To achieve the objective of the study the following work plan and methodologies were applied:

- Sequencing of the earliest identified *mcr-1* positive *S.* Paratyphi B *d*Ta+ ST28 isolate from Germany using PacBio technology for reference strain definition
- Whole-genome sequencing using MiSeq technology of *mcr-1* positive and selected *mcr-1* negative *S.* Paratyphi B *d*Ta+ ST28 isolates from Germany (99 isolates) and Denmark (17 isolates) received between 2008 to 2016 representing different multidrug-resistance patterns and covering sources from food-producing animals and food products thereof
- Determination of the phylogeny and antimicrobial resistance determinants of isolates
- Determination of the diversity of *mcr-1* positive plasmids and comparison of plasmid sequences

Methods and results

Approximately 400 *S.* Paratyphi B *d*Ta+ isolates originating from food producing animals and food products, received from 2006 to 2016 in the German National Reference Laboratory for Salmonella were selected for PCR screening for the presence of the *mcr-1* gene. Altogether 63 *mcr-1* positive *S.* Paratyphi B *d*Ta+ ST28 isolates were subjected to WGS. Sequencing data assembled and was analysed using the Bacterial Analysis Pipeline provided by the Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/cge/) and SNP analysis was performed using BioNumerics 7.6 (Applied Maths). Results showed that the *mcr-1* gene was located on plasmids belonging to the IncHI2 or IncX4 replicon group. Strains isolated from 2008 to 2011 tend to carry the *mcr-1* gene on large multidrug-resistant IncHI2 plasmids and build a cluster in the phylogenetic tree, whereas strains isolated after 2011 mainly carry *mcr-1* on IncX4 plasmids and show a higher phylogenetic diversity. Some of the *S.* Paratyphi B *d*TA+ strains from Denmark were found to cluster together with German isolates.

Additional notes

The project will be finished within the year 2018, beyond the end of the ENGAGE period. Preliminary results have been presented on the ENGAGE workshop in 2016 (Warsaw, Poland) and the interim meeting in 2017 (Parma, Italy). Furthermore, a talk has been hold at the ECCMID 2017 conference in Vienna.

A first publication appeared in ASM Genome Announcement describing the complete genome sequence of the earliest identified *mcr-1* positive *Salmonella* Paratyphi B *d*Ta+ strain in the collection of the National Reference Laboratory for Salmonella in Germany:

Borowiak M, Hammerl JA, Fischer J, Szabo I, Malorny B. 2017. Complete genome sequence of *Salmonella enterica* subsp. *enterica* serovar Paratyphi B sequence type 28 harboring *mcr-1*. Genome Announc 5:e00991-17

Appendix K.3

Project title	Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B
Consortium partner	BfR
Scientific lead	Burkhard Malorny, Maria Borowiak
Other people involved	Rene Hendriksen

Purpose

The purpose of the study was to identify a novel colistin resistance meachanism in *mcr-1*, *mcr-2*, *mcr-3* and *mcr-4* negative *S.* Paratyphi B *d*Ta+ isolates with colistin MIC values > 2 mg/L which were received in the years from 2011 to 2016.

Background

Liu and colleagues (2015) have recently described a mobilizable colistin resistance gene, *mcr-1,* located on a plasmid. Among the reported *mcr-1* positive Enterobacteriaceae isolates of *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B (*S*. Paratyphi B *d*Ta+), formally called *S*. Paratyphi B variant Java, are described. In Germany, Belgium and the Netherlands a multidrug-resistant linage of *S*. Paratyphi B *d*Ta+ belonging to the sequence type ST-28 persists in the poultry production since the 1990s. Between 2011 and 2016, 86 (21 %) of 414 tested *S*. Paraytphi B isolates from the strain collection of the German National Reference Laboratory for the Analysis and Testing of Zoonoses (NRL Salmonella) were determined to have a non-wild-type phenotype for colistin (MIC > 2 mg/L). PCR screening revealed, that 54 isolates carried the *mcr-1* and none of them the *mcr-2, mcr-3* and *mcr-4* gene. Of the remaining 32 strains, twelve isolates showed non-wild-type phenotypes for ampicillin, ciprofloxacin, colistin, nalidixic acid, sulphamethoxazole, tetracycline, trimethoprim and in some cases tigecycline. This phenotypic pattern was unique to strains with unknown colistin resistance mechanism.

Objectives

Before this study, plasmid-mediated mobilised colistin resistance was known to be caused by phosphoethanolamine transferases termed MCR-1, MCR-2, MCR-3 and MCR-4. However, this study focussed on the dissection of a novel resistance mechanism in *mcr-1*, *mcr-2*, *mcr-3* and *mcr-4* negative *S*. Paratyphi B *d*Ta+ isolates with colistin MIC values > 2 mg/L.

Project description

To achieve the objective of the study the following work plan and methodologies were applied:

- Sequencing of a randomly selected isolate with non-wild-type phenotypes for ampicillin, ciprofloxacin, colistin, nalidixic acid, sulphamethoxazole, tetracycline and trimethoprim (this phenotypic pattern was unique to strains with unknown colistin resistance mechanism) using Illumina MiSeq technology
- Bioinformatics analysis of obtained sequencing data including assembly and RASTtk annotation for identification of novel phosphoethanolamine transferase genes
- PCR-screening of all 32 colistin resistant S. Paratyphi B *d*Ta+ isolates with unknown colistin resistance mechanism for the presence of the newly identified gene
- Functional characterisation of the novel gene

Methods and results

A selected isolate of *mcr-1*, *mcr-2*, *mcr-3 and mcr-4* negative *S*. Paratyphi B *d*Ta+ isolates with colistin MIC values > 2 mg/L from the strain collection of the German National Reference Laboratory for Salmonella was investigated by WGS and bioinformatics analysis (RASTtk annotation) as described in Borowiack et al., 2017 (see below). Analysis of sequence data lead to the discovery of a novel transposon-associated and plasmid-encoded phosphoethanolamine transferase involved in colistin resistance. The respective gene, further termed as *mcr-5* (1644 bp), was part of a 7337 bp transposon of the Tn*3*-family and located on related multi-copy ColE-type plasmids. Subsequently PCR screening, S1-PFGE and DNA-DNA hybridisation were performed to analyse the prevalence and localisation of the *mcr-5* gene. Altogether 14 isolates form poultry and food thereof collected between 2011 and 2013 were found positive for *mcr-5*. Interestingly, in one isolate an additional subclone with a chromosomal location of the *mcr-5* transposon was observed. All isolates harboured similar plasmids. Cloning and transformation experiments in *Escherichia coli* DH5α and *S*. Paratyphi B *d*Ta+ control strains were carried out and the activity of MCR-5 was confirmed *in vitro* by MIC testing.

Additional notes

The project has been finished and is published:

Borowiak M, Fischer J, Hammerl JA, Hendriksen RS, Szabo I, Malorny B: Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B. J Antimicrob Chemother. 2017 72(12):3317-3324

Appendix K.4

Project title	VIM-1 producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany
Consortium partner	BfR
Scientific lead	Burkhard Malorny, Maria Borowiak
Other people involved	-

Purpose

The purpose of this study was the characterization of bla_{VIM}-harbouring *S.* Infantis recovered from swine and minced pork meat in Germany in 2015 and 2016.

Background

Carbapenems are considered as last-line clinical antibiotics used to treat severe human infections caused by multidrug-resistant Gram-negative bacteria. In the last years, carbapenemase-producing Enterobacteriaceae (CPEs) leading to carbapenem resistance or carbapenem non-susceptibility spread in human settings worldwide and are associated with major public health concerns. First detections of VIM-1 carbapenemase-producing *S.* Infantis and *E. coli* isolates in livestock farms in Germany in 2011 raised concern on their real prevalence in animals and their potential to be transferred to human settings. Indeed, *Salmonella* is a well-known zoonotic pathogen commonly transferred via contaminated food-products. Among them, *S.* Infantis is one of the leading causes of human salmonellosis in Europe.

In the frame of the routine diagnostic, monitoring and control programs, the National Reference Laboratory for Salmonella in Germany received 145 *S.* Infantis isolates in 2015 (33.8% from primary production; 38.6% from food; 13.1% from feed; 14.5% from other sources). In the routine antimicrobial susceptibility testing one of these *S.* Infantis strains, isolated from minced pork meat produced in Germany (15-SA01028) showed a microbiological resistance ("non-wild-type" phenotype) against meropenem (MIC=0.5 mg/L), imipenem (MIC=4 mg/L) and ertapenem (MIC=0.12 mg/L) This isolate, together with another *S.* Infantis strain isolated in 2016 from a sick piglet (16-SA00749) which showed decreased susceptibility to meropenem (MIC=0.12 mg/L) were analysed by whole genome sequencing.

Objectives

The objective of this study was to characterize two VIM-1 positive *S.* Infantis strains isolated in 2015 from minced pork meat produced in Germany (15-SA01028) and 2016 from a sick piglet (16-SA00749) by whole genome sequencing.

Project description

To achieve the objective of the study the following work plan and methodologies have been applied:

- Susceptibility testing in concordance with the European Commission Implementing Decision 2013/652/EU) and following EUCAST epidemiological cut-off values (ECOFFs, http://www.eucast.org)
- Comparison of newly detected VIM-1 positive isolates with previously identified isolates by S1 nuclease pulsed-field gel electrophoresis (S1-PFGE)
- Sequencing of newly detected VIM-1 positive isolates using Illumina MiSeq technology and bioinformatics analysis of obtained sequencing data

Methods and results

Two *S*. Infantis isolates with microbiological resistance (15-SA01028) or decreased susceptibility (16-SA00749) to meropenem were investigated by WGS and bioinformatics analysis as described in Borowiak et al., 2017 and 2018 (see below). Both isolates were found to be positive for VIM-1 carbapenemases. S1-PFGE and mapping of NGS data to already published resistance plasmid sequences revealed that the plasmid harboured by the food isolate (15-SA01028) is 100% similar to the previously published plasmid sequence of isolate R27 (pRH-R27) from 2011 whereas the isolate from swine (16-SA00749) harboured a plasmid with 96% similarity to pRH-R27. These findings hint towards a link between the isolates and to a transmission of this plasmid or these *Salmonella* isolates from the primary production into the food chain. The occurrence of carbapenemase-producing Enterobacteria (CPE) in food and food-producing animals might bear a risk of getting colonized with CPEs and raises major public health concerns.

Additional notes

The project has been finished and is published:

Borowiak M, Szabo I, Baumann B, Junker E, Hammerl JA, Kaesbohrer A, Malorny B, Fischer J: VIM-1-producing *Salmonella* Infantis isolated from swine and minced pork meat in Germany. J Antimicrob Chemother. 2017 Jul 1;72(7):2131-2133

Borowiak M, Fischer J, Baumann B, Hammerl JA, Szabo I, Malorny B. 2018. Complete genome sequence of a VIM-1-producing *Salmonella enterica* subsp. *enterica* serovar Infantis isolate derived from minced pork meat. Genome Announc 6:e00327-18

Appendix K.5

Project title	*Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and antimicrobial resistance (AMR) genes
Consortium partners	IZSLT (lead), DTU
Scientific lead	Antonio Battisti, Rene Hendriksen
Other people involved	Alessia Franco, Patricia Alba, Pimlapas Leekitcharoenphon, Susanne Karlsmose Pedersen

Purpose

The purpose of the study was to provide molecular data based on WGS analysis in order to investigate similarities and differences within *Salmonella enterica* subsp. *enterica* serovar Infantis (*S.* Infantis) across Europe and across animal and food sources. This project was intended as a contribution to facilitating source attribution assessment of human cases arising in the European Union (EU), also when isolates investigated may not be outbreak-associated or whose epidemic nature has gone unrecognised. A collaborative EU study seemed to be the ideal approach for the purpose.

Background

Salmonella Infantis is emerging worldwide. It is the most frequently reported serovar in broilers (45.6%), the most prominent one in broiler meat (47.4%) and the fourth most prevalent one in non-typhoidal *Salmonella* (NTS) human infections in Europe (EFSA EU Summary Report 2016). Over the last decade, MDR *S*. Infantis has increasingly been reported in Italy from food-animals and humans, and is also highly prevalent in the broiler meat industry, in several European countries (EFSA AMR EU Summary Report 2016).

Additionally, according to the Italian antimicrobial resistance (AMR) monitoring data on isolates from the National Control Program (NCP) in broiler chicken flocks in 2012, 2013, and 2014, Extended-Spectrum Cephalosporin (ESC)-resistance (R) rates in *Salmonella* spp., of which most were *S*. Infantis, have increased, from 15.6% (12/77, 95% CI 8.3–25.6%) in 2012 to 27.27% (18/66, 95% CI 17.03-39.64%) in 2014, reaching in 2016 moderate level of resistance (12%).

In cross-sectional studies performed in Italian broiler sector at slaughter in 2014 and 2016 (sampling frame: Commission Implementing Decision 2013/652/EU), *S.* Infantis accounted for 75% and 90% of all isolates detected, respectively, with an among flock prevalence of 9.6% (68/709) in 2014 and 8.7% (70/807) in 2016. An emerging clone harbouring a megaplasmid (around 300 kb) termed pESI, which carries virulence, fitness and MDR genes/traits, along with CTX-M-1 ESBL in an increasing proportion of isolates, was detected in these surveys (3/90 in 2014 and 16/77 in 2016 were MDR, ESBL-producing S. Infantis, respectively) (EFSA AMR EU Summary Report 2016).

Objectives

Whole Genome Sequencing was the method of choice in this project to assess both population structure and phylogenetic relationships among isolates, and to investigate on genetic basis for

virulence, antimicrobial resistance, presence and characterisation of plasmids with the aim of comparing isolates from different stages of the food chain, in order to provide input data for epidemiology and risk assessment purposes.

The objectives of the study were to characterize *S.* Infantis isolates from EU animal primary productions comprehensively by Whole Genome Sequencing, and to provide genomic insight to better elucidate *transmission pathways* along the food chain; from primary production systems (e. g. poultry, pigs), foodstuffs thereof (e. g. meats) to human cases arising in the community.

Project description

A collection of *S.* Infantis isolates from different primary productions and foodstuffs of animal origin, was investigated by *Whole Genome Sequencing* (WGS) and compared with isolates available from human cases.

Selection criteria for isolates: Isolates from population-based studies (e.g. National Control Programmes, monitoring activities at farm or at retail level (food) were preferred. However, in order to account for as much of geographical origins as possible at EU level, isolates from convenience sampling activities in some Member States (MS) were also investigated. Human isolates were selected among those obtained by surveillance activities of cases occurring in the community. An outgroup of isolates, detected outside the EU, were included mainly on a convenience basis criteria, albeit accounting for different geographical origins.

Sample size of isolates under study had taken into account prevalence/isolation rates from different sources, full susceptibility or resistance to selected antimicrobial classes (e.g. CIAs) and multidrug-resistance (MDR) patterns and availability across EU.

In order to gather as much variability as possible across EU for the purposes of the study, it was advisable to increase the minimum sample size provisionally set for the dominant *Salmonella enterica* serovars, as described within the ENGAGE scopes and agreement among the Consortium partners and EFSA.

Since recent research studies (Franco et al., 2015) indicated that in Italy both ESC-S and ESC-R isolates are mainly transmitted by the broiler chicken poultry system to humans, with rare isolates of pig origin also described, these two sources, and any relevant/additional animal/food sources, were further investigated at EU level.

A minimum samples size per EU Member State was set on each stage/origin (animal, food, human), with additional isolates requested to MS laboratories for selected sources or AMR patterns or accessory genome content, when necessary.

Isolates selected: On the basis of the above-mentioned criteria, a total of N=229 *S.* Infantis isolated from humans (N=64), from animal (N=115) and food/environment (N=50) sources in the frame of National Control Programmes and monitoring activities were selected from 5 countries: 150 in Italy, 20 in Ireland, 18 in Luxembourg, 26 in Netherlands and 15 in Poland. In addition, 34 sequenced genomes were provided by APHA (UK), 38 by BfR (Germany) and 80 by EURL-AR DTU-Food (Denmark). Additionally, seven pESI-like positive S. Infantis genomes from USA have been downloaded from ENA public repository (Tate et al., 2017).

Methods and Results

All isolates were whole genome sequenced by the IZSLT or by other ENGAGE partners using Illumina technology. Raw reads were assembled using SPAdes v3.11.0 after a quality check performed with FastQC v0.11.5 and Trimmomatic v0.11.5.

The obtained sequences were analysed by using different bioinformatic tools available at the Center for Genomic Epidemiology http://cge.cbs.dtu.dk/services/all.php or with Blast v2.2.31 using the CGE databases as reference, for the following purposes:

- Detection of antimicrobial resistance genes (both horizontally acquired and known chromosomal point mutations), selected virulence genes, MLST, pMLST, and genetic environment of selected resistances (e.g. ESC-R, Plasmid-Mediated Quinolone-R (PMQR), Colistin resistance (COL-R) (whenever detected).

- Construction of a SNP-based phylogeny for phylogenetic and phylogeographic insight into transmission patterns between animal primary productions and humans.

- Identification and characterization of selected plasmids (e. g those carrying selected virulence genes or AMR pattern), also comparing them across productions and EU countries.

At present, all the raw reads have been assembled and analysed for the presence of specific genetic markers associated with the pESI-like megaplasmid described in Italy in 2015 (Franco et al., 2015), the plasmid content and accessory antibiotic resistance genes. Remarkably, specific markers of the pESI-like megaplasmid, such as pESI backbone, *fim* and K88 genes have been found in isolates from all the countries participating in the study. Further *in silico* analysis to confirm the presence of *S.* Infantis isolates harbouring the pESI-like megaplasmid are still ongoing.

WGS sequences obtained from isolates from Italy, Ireland, Luxembourg, Netherlands, Poland, UK and Denmark have been submitted to ResFinder 3.0 to investigate the presence of specific chromosomal point mutations related to fluoroquinolones resistance.

The SNPs-based phylogenetic and phylogeographic analysis by using the CSI phylogeny is still ongoing. At present, a draft phylogeny tree with a subset (N= 265) of the sequenced isolates from different sources retrieved from Italy, Poland, Ireland, UK, Denmark, Finland and USA, has been constructed. Preliminary results suggest that no geographical links among the different *S.* Infantis clones have been detected so far.

Additional notes

- To date, the IZSLT, in collaboration with the DTU, are doing further analysis on accessory genome (especially pESI plasmid) harboured by the isolates in the collection. Most likely, a manuscript that includes part of the results of this proof of concept and with a working title of "*Salmonella* Infantis in Italy and EU: phylogeny and plasmid carrying virulence, fitness and AMR genes" will be ready to be submitted in a peer-reviewed scientific journal by the end of summer 2018.

Preliminary results were presented in the following meetings and workshops:

-Battisti A., Franco A. (NRL-AR Italy). *Salmonella* Infantis clones and emerging ESC-R in Italy: differences and similarities of strains and plasmids in Europe and USA. Scientific Network for Zoonoses Monitoring Data, 7th specific meeting on Antimicrobial Resistance data reporting, 08-09 November 2017, Parma, Italy,

https://www.efsa.europa.eu/en/events/event/171108;
https://www.efsa.europa.eu/sites/default/files/event/171108-0-m.pdf

-Battisti A. (NRL-AR Italy). Spread of an emerging clone of MDR, ESBL-producing *Salmonella* Infantis harbouring a conjugative megaplasmid in Italy. 3rd joint meeting on AMR in *Salmonella* and Campylobacter, FWD-Network and EURL-AR Network. 6-7 April 2017, Copenhagen, Denmark

https://ecdc.europa.eu/en/news-events/third-joint-workshop-amr-salmonella-and-campylobacter

-Battisti A. (NRL-AR Italy). Spread of an emerging clone of MDR, ESBL-producing *Salmonella* Infantis harbouring a conjugative megaplasmid in Italy. Workshop EURL-AR 2016, Kgs. Lyngby April/2016

https://www.eurl-ar.eu/CustomerData/Files/Folders/3-workshop-kgs-lyngby-april2016/10_spread-of-an-emerging-clone-of-mdr-esbl-producing-salmonella-infantis-harbouring-a-conjug.pdf

References

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control), 2017. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. EFSA Journal 2017;15(12):5077, 228 pp. doi:10.2903/j.efsa.2017.5077

Franco A, Leekitcharoenphon P, Feltrin F, Alba P, Cordaro G, Iurescia M, Tolli R, D'Incau M, Staffolani M, Di Giannatale E, Hendriksen RS and Battisti A, 2015. Emergence of a Clonal Lineage of Multidrug-Resistant ESBL-Producing Salmonella Infantis Transmitted from Broilers and Broiler Meat to Humans in Italy between 2011 and 2014. PLoS One, 10(12):e0144802. doi: 10.1371/journal.pone.0144802.

Tate H, Folster JP, Hsu CH, Chen J, Hoffmann M, Li C, Morales C, Tyson GH, Mukherjee S, Brown AC, Green A, Wilson W, Dessai U, Abbott J, Joseph L, Haro J, Ayers S, McDermott PF and Zhao S, 2017. Comparative Analysis of Extended-Spectrum-β-Lactamase CTX-M-65-Producing Salmonella enterica Serovar Infantis Isolates from Humans, Food Animals, and Retail Chickens in the United States. Antimicrobial Agents and Chemotherapy, 61:e488-17. doi: 10.1128/AAC.00488-17

Appendix K.6

Project title	Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* in Italy
Consortium partners	IZSLT (lead) and DTU
Scientific lead	Antonio Battisti, Rene Hendriksen
Other people involved	Alessia Franco, Patricia Alba, Pimlapas Leekitcharoenphon, Valeria Bortolaia, Virginia Carfora

Purpose

The purpose of this study was to investigate the epidemiology of transferable colistin resistance mediated by *mcr* genes in food producing animals in Italy.

Background

Since the description of the *mcr*-1 gene on a transferable plasmid encoding resistance to colistin from *E. coli in* China in 2015 (Liu et al., 2016), at least 32 countries from the five continents have found *mcr*-1 in *Enterobacteriaceae* isolates from different sources including humans, animals and foodstuffs (Xavier et al., 2016). In the meanwhile, 11 *mcr*-1 variants have been described and four new *mcr* genes, namely *mcr*-2, *mcr*-3, *mcr*-4 and *mcr*-5 and its variants, have been found.

In Europe, the presence of *mcr*-1 was first detected in *E. coli* from poultry meat and humans in Denmark (Hasman et al., 2015). Subsequently, this gene was detected in *Enterobacteriaceae* isolated from different sources in almost all European countries, including Italy (Battisti, 2016).

The epidemiology of transferable *mcr*-mediated colistin resistance is evolving rapidly and in the next future may hamper effective antimicrobial therapeutic approach to invasive infections caused by multidrug-resistant Gram-negative bacteria at both hospital and community level. Thus, timely information on prevalence, trends and variants of *mcr*-positive isolates is needed to enhance surveillance, and implement prevention and control measures.

Objectives

The aim of this study was to determine the prevalence of colistin resistance, the molecular epidemiology of *mcr*-mediated colistin resistance genes and their genetic environment in commensal *E. coli, Extended Spectrum Beta-Lactamase (*ESBL)/AmpC-producing *E. coli*, and *Salmonella* spp. in different primary productions and foodstuffs of animal origin in Italy, over the last three-years (2014-2016).

Project description

A collection of commensal *E. coli,* Extended Spectrum Beta-Lactamase (ESBL)/AmpC-producing *E. coli*, and *Salmonella* spp. isolates from Italian population-based studies (e. g. National Control Programmes,

monitoring activities at farm or at retail level (food)), conducted in a three-year period (2014-2016), was investigated by *Whole Genome Sequencing* (WGS).

Isolates selected: A total of 55 *E. coli* and 14 *Salmonella* of different serotypes, phenotypically resistant to colistin (COL-R; *MIC value ≥ 4 mg/L*), with the majority of them being also MDR including Extended-Spectrum Cephalosporin-resistant (ESC-R), along with 6 fully susceptible *E. coli* isolates included as negative control, were selected. All isolates had been collected by the Italian National Reference Laboratory for Antimicrobial Resistance (NRL-AR) in the frame of National Control Programmes and monitoring activities (2014-2015-2016) from different primary productions (fattening turkey, broiler chicken, fattening pigs and calves).

After the partial analysis of the results, additional three multidrug resistant (MDR) *S*. Infantis displaying a colistin MIC value ≥ 4 mg/L and *mcr*-1 positive by PCR, with two of them being also extended-spectrum cephalosporin-resistant (ESC-R), collected in the frame of National Control Programmes and monitoring activities (2016-2017) were sequenced after the end of the ENGAGE project (December 2017) and then included in this project.

For a better interpretation of the results, the isolates were divided in sub-groups according to the species, the serotype and the source. At this regard, at least 2 subgroups were deeply investigated:

- 42 commensal and ESBL/AmpC-producing *E. coli* and *Salmonella* spp. phenotypically resistant to colistin and *mcr* positive by PCR, isolated in the frame of the EU harmonised antimicrobial resistance (AMR) monitoring conducted in poultry (2014) and fattening pigs or calves (2015). Regarding their origin, 31 isolates were collected from turkeys (28 *E. coli* and 3 *Salmonella enterica* isolates), 5 *E. coli* originated from pig samples and 6 *E. coli* from bovine samples.

- 4 multidrug resistant (MDR) *S*. Infantis from broilers (N=3) and broiler meat samples (N=1), displaying a colistin MIC value ≥ 4 mg/L and *mcr*-1 positive by PCR with two of them being also extended-spectrum cephalosporin-resistant (ESC-R), collected in the frame of National Control Programmes and monitoring activities (2016-2017).

Methods and Results

The sequences obtained from all isolates were analysed as follows:

- Assembly of the raw reads using the pipeline of the Center for Genomic Epidemiology (CGE). In some particular circumstances, as in the case of the 4 MDR *S.* Infantis *mcr*-1 positive isolates, the sequences were analysed as reported in Appendix K.5.

- Identification of serotypes, Sequence Types (STs), presence of virulence and accessory resistance genes by using suitable bioinformatic tools, as the CGE pipeline.

- Determination of the presence of plasmids in the isolates, through detection of the different plasmid replicon types in the bioinformatic analysis output.

- Description of the *mcr* variants harboured by the subset of the isolates sequenced, comparing them with the nucleotide archive of the NCBI using BLAST.

- Description and characterization of the *mcr* gene molecular environment, obtained by the identification and annotation of the genes located upstream and downstream of *mcr* in the assembled sequences.

- When appropriated and based on previously collected epidemiological and molecular data, a SNPs based phylogeny of selected isolates compared with historical isolates from the italian NRL collection was built.

The main findings derived from this study were:

- The *E. coli* population harbouring colistin resistance mediated by *mcr* was highly heterogeneous, as proved by the diversity of STs found and by the variety of virulence and resistance genes detected.

- The commensal and the ESBL/AmpC-producing colistin resistant *E. coli* populations harboured a high variability of *mcr* genes. In total, three different *mcr* genes were detected (*mcr*-1, *mcr*-2 and *mcr*-3) and six variants, including one variant not previously described (*mcr*-1.1, *mcr*-1.2, *mcr*-1.13, *mcr*-3.2, *mcr*-4.2, *mcr*-4.3).

- The *mcr* gene was detected in different *Salmonella* serovars circulating in Italy, including *S*. Typhimurium, *S*. Newport, *S*. Blockley and *S*. Infantis, isolated from primary productions or food derived from animals.

- The *S*. Infantis isolates contained both pESI-like megaplasmids and IncX4 plasmids harbouring *mcr*-1.1 and also belonged to the emerging, pESI-like positive, ESBL-producing clone described in Italy in 2015 (Franco et al., 2015).

Additional notes

-Part of this study has been included in a scientific paper:

Alba P, Leekitcharoenphon P, Franco A, Feltrin F, Ianzano A, Caprioli A, Stravino F, Hendriksen R, Bortolaia V, Battisti A. Molecular epidemiology of *mcr*-encoded colistin resistance in *Enterobacteriaceae* from food-producing animals in Italy revealed through the EU harmonised antimicrobial resistance monitoring. Accepted for publication in the peer-reviewed scientific journal *Frontiers in Microbiology* (May 2018). doi: 10.3389/fmicb.2018.01217.

-To date, the IZSLT in collaboration with the DTU are preparing a manuscript entitled "Colistin resistance mediated by *mcr*-1 in ESBL-producing, multidrug-resistant *Salmonella* Infantis in broiler chicken industry, Italy (2016-2017)" to be submitted in a peer-reviewed scientific journal, that includes part of the results of this proof of concept.

Part of the results has been presented in the following international conferences:

- Alba P, Feltrin F, Iurescia M, Amoruso R, Donati V, Caprioli A, Leekitcharoenphon P, Hendriksen RS, Battisti A, Franco A. Transferable colistin resistance mediated by the *mcr*- 1 gene is widespread among *Escherichia coli* and is emerging in *Salmonella* in the Italian fattening turkey industry. 26th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID 2016). 9-12 April, Amsterdam, Netherlands.

- Alba P., Feltrin F., Iurescia M., Amoruso R., Donati V., Caprioli A., Leekitcharoenphon P., Hendriksen R., Franco A., Battisti A. Transferable colistin resistance mediated by the *mcr*-1 gene is widespread among *Escherichia coli* and is emerging in *Salmonella* in the italian fattening turkey industry. 2018. 18th International Symposium of the World Association of Veterinary Laboratory Diagnosticians. Sorrento (Italy).

References

Battisti A, 2014. Antibiotic resistance - Italy: colistin, *mcr*-1, *E. coli*, turkeys, 2014. ProMED-mail post. 2016. Archive Number: 20160113.3933461.

Hasman H, Hammerum AM, Hansen F, Hendriksen RS, Olesen B, Agersø Y, Zankari E, Leekitcharoenphon P, Stegger M, Kaas RS, Cavaco LM, Hansen DS, Aarestrup FM and Skov RL, 2015. Detection of *mcr*-1 encoding plasmid-mediated colistin-resistant isolates from human bloodstream infection and imported chicken meat, Denmark 2015. *Euro Surveillance,* 20. doi: 10.2807/1560-7917.ES.2015.20.49.30085.

Franco A, Leekitcharoenphon P, Feltrin F, Alba P, Cordaro G, Iurescia M, Tolli R, D'Incau M, Staffolani M, Di Giannatale E, Hendriksen RS and Battisti A, 2015. Emergence of a Clonal Lineage of Multidrug-Resistant ESBL-Producing *Salmonella* Infantis Transmitted from Broilers and Broiler Meat to Humans in Italy between 2011 and 2014. PLoS One, 10(12):e0144802. doi: 10.1371/journal.pone.0144802.

Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu LF, Gu D4, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu JH and Shen J, 2016. Emergence of plasmid-mediated colistin resistance mechanism *mcr*-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infectious Diseases, 16:161-168. doi: 10.1016/S1473-3099(15)00424-7.

Xavier BB, Lammens C, Ruhal R, Kumar-Singh S, Butaye P, Goossens H and Malhotra-Kumar S, 2016. Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli*, Belgium, June 2016. *Euro Surveillance,* 21(27). doi: 10.2807/1560-7917.ES.2016.21.27.30280.

Appendix K.7

Project title	Genomic diversity of *Salmonella* Derby in different European countries
Consortium partner	IZSVe
Scientific lead	Antonia Ricci
Other people involved	Eleonora Mastrorilli, Sara Petrin, Alessandra Longo, Carmen Losasso, Lisa Barco

Purpose

To characterise through WGS a collection of isolates of *S*. Derby from different sources, different steps of the food chain and representative of isolates circulating in the geographical area of competence of ENGAGE partners.

Background

S. Derby continues to represent an important serovar causing human infection in Europe (EU). In 2016, according to the EFSA report on zoonoses, S. Derby was the fifth most common serovar notified from human domestic cases of salmonellosis in EU (EFSA and ECDC, 2017) confirming the situation of the previous years.

S. Derby is generally associated to pig chain, and the strong association of the serovar with this source is evident for all the EU countries. However, with a lower frequency, *S.* Derby has been notified also from other sources, such as poultry and in particular turkeys. Few countries reported the isolation of *S*. Derby from sources different from pig during the last couple of years. In particular, UK and Ireland reported a notable number of *S*. Derby isolates from turkey flocks. Since the notification of this serovar is not mandatory (according to Reg. (EC) No 2160/2003 on the control of "*Salmonella* and other specified food-borne zoonotic agents" and subsequent implementing acts), it is not clear if the isolation of *S.* Derby from sources different form pigs is a local problem for some specific countries or if this finding is simply related to the different approaches followed by Member States in reporting *Salmonella* data to EFSA.

Previous studies in UK have demonstrated that a pathogenicity island discovered in the genome of *S.* Derby (SPI-23) plays an important role in its adaptation to porcine jejunum over porcine colon. Moreover, through comparative genomic, it has been demonstrated that *S.* Derby isolates from pig and turkey chains belong to different clonal lineages (Hayward *et al.,* 2016), confirming that different sources are associated to different clones of *S*. Derby.

Looking at the persistence of *S*. Derby along the food chain there is the evidence that swine isolates are generally obtained from animals as well as from food sources. Conversely, considering isolates from other species, such as turkeys, the isolation is most frequent from animals than from foodstuffs. This statement is based on the data collected in Italy, but it has been confirmed also at EU level in the context of analysis of *Salmonella* serovars reported to EFSA for the 2015 and 2016 annual summary report on zoonoses (EFSA and ECDC, 2016, 2017).

Objectives

The aim of this project was the characterisation of *S.* Derby isolates from different sources and lineages isolated from different countries, with the final goal of identifying the clones that circulate at the EU level. Moreover, we wanted to identify specific genetic features that could explain host adaptation and persistence of *S.* Derby isolates along the food chain in relation to specific sources, in order to ascertain if any differences occur among S. Derby isolated from different sources at different points of the food chain. Also human isolates were included in the investigation in order to infer the main sources of human infection associated to this serovar and the most virulent lineages.

Project description

Partners were asked to contribute to the project with 60-80 isolates from pigs, turkeys, humans and other species; isolates from both animals and foodstuffs were considered, but participants were asked to contribute also with human isolates.

The project gained the favour of all the ENGAGE partners, who contributed with their own isolates/sequences. A total of 342 sequences of *S.* Derby were retrieved from 7 partners (IZSVe, BfR, DTU, NVRI, APHA, PHE and NIPH-NIH) whilst additional 87 strains were internally provided from PHE, for a grand total of 429 strains. This huge participation confirmed the interest toward the epidemiological issues related to this serovar.

Methods and results

Quality metrics checking, species identification, MLST based on the 7 housekeeping genes, MLST-guided serotyping, prediction/serotyping from raw sequencing reads, AMR in-silico characterization, eBURST Group (eBG) assignation according to Achtman *et al.* and single nucleotide polymorphism (SNP) address assignation were performed using the internally developed PHE pipeline (including KmerID (https://github.com/phe-bioinformatics/kmerid), MOST (Tewolde et al. 2016, PeerJ - https://github.com/phe-bioinformatics/MOST), PHEnix (http://phenix.readthedocs.io/en/latest/) and SnapperDB (Dallman et al., 2018 Bioinformatics https://github.com/phe-bioinformatics/snapperdb)). Isolates belonging to different eBGS were further investigated by multiple alignment in order to ascertain their level of similarity in relation to sources, geographical area of origin and other distinctive features. A selection of isolates were subjected to *in-silico Salmonella* pathogenicity island (SPI) detection to characterize them and to ascertain the presence of SPI-23 to test the hypothesis of its role in characterizing host adaptability.

With the end of the project both the sequencing of internal isolates and the collection of sequences by all the partners was finalized. A preliminary analysis of the collected sequences was carried out in the context of the twinning between IZSVe and PHE that took place in summer 2017. A subset of about 200 isolates, available when the twinning took place, were analysed according to the standard PHE internal pipeline. The analysed isolates belonged to 9 different ST (39,40, 71, 682, 683, 1326, 3135, 3857 and 3871), corresponding to three eBGs (57, 244 and 264). Two eBGs, containing more than one samples, were further investigated: all samples belonging to eBG_264 belonged to the same 50 SNP difference cluster except one sample; samples sharing the highest degree of similarity were isolated in the same country (Italy of Germany); all samples belonging to eBG_57 belonged to two main clusters, depending on their ST (39 or 40). *In-silico Salmonella* pathogenicity island (SPI) detection confirmed the massive presence of a panel of different SPIs: SPI1, SPI9 and C63Pib; SPI5 were found in all but one sample; SPI2, SPI4 were seen in more than 80 samples.

Additional notes

The conclusion of the analyses is planned by 2018, beyond the end of the ENAGE project period.

The analytical pipeline previously described will be extended to the entire collection of sequences. Moreover, the preliminary analyses carried out confirmed the need of getting a clearer picture of the genetic markers associated to *Salmonella* virulence. Thus, IZSVe decided to perform a systematic review aimed at identifying the genetic features which are involved in *Salmonella* virulence, with particular interest to genetic markers associated to *S.* Derby, in order to enlarge the targets of analyses for the entire panel of the sequences collected.

This project has been a great opportunity of collaboration among ENGAGE partners in terms of sequences sharing as well as an opportunity for the proposing partner (IZSVe) to be trained by the PHE on the analytical approaches used on their routine.

References

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control), 2016. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. EFSA Journal 2016;14(12):4634, 231 pp. doi:10.2903/j.efsa.2016.4634

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control), 2017. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. EFSA Journal 2017;15(12):5077, 228 pp. doi:10.2903/j.efsa.2017.5077

Hayward MR, Petrovska L, Jansen VA and Woodward MJ, 2016. Population structure and associated phenotypes of *Salmonella enterica* serovars Derby and Mbandaka overlap with host range. BMC Microbiology, 16:15. doi: 10.1186/s12866-016-0628-4.

Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG. Hale JL, Harbottle H, Uesbeck A, dougan G, Harrison LH, Brisse S and the *S. enterica* MLST Study Group, 2012. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS pathogens, 8(6), e1002776.

Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J and Underwood A, 2016. MOST: a modified MLST typing tool based on short read sequencing. PeerJ, 4, e2308. doi: 10.7717/peerj.2308

Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R, Underwood A, Jenkins C and Grant K, 2018. SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. Bioinformatics, doi: 10.1093/bioinformatics/bty212

Appendix K.8

Project title	Whole genome characterization of *Salmonella* Napoli isolates spanning 2005-2015: a national issue of international interest
Consortium partner	IZSVe
Scientific lead	Antonia Ricci
Other people involved	Eleonora Mastrorilli, Sara Petrin, Alessandra Longo, Carmen Losasso, Lisa Barco

Purpose

To characterise through WGS a collection of *S.* Napoli isolates to increase the knowledge on the ecology of this serovar and the main risk factors for human infections.

Background

S. Napoli is among the top serovars causing human infections in Italy and accounted for 5.9% of salmonellosis cases reported in this country during the last two years (Sabatucci et al, 2018). The relevance of this serovar is demonstrated by the increasing number of cases notified each year: during 2000–2006, an increase of 140% of *S*. Napoli cases was reported in Europe, mostly (87%) related to Italy, France and Switzerland (Huedo et al., 2017; Sabatucci et al., 2018).

This serovar is relatively uncommon in other European countries. Several outbreaks, caused by *S*. Napoli, have been documented during the last years (1982-2015) (Sabatucci et al., 2018). In these cases the sources of outbreaks were related to exported Italian food products, mostly vegetabes.

S. Napoli are generally isolated from humans and environment, whereas strains from animals and food are quite rare. Several studies, using different approaches, tried to infer the sources of infection for this serovar; however the reservoirs and transmission pathways are still partly unknown (Fisher et al., 2009; Graziani et al., 2011; Graziani et al., 2015). Moreover, the interest toward this serovar is triggered also by epidemiological, clinical and molecular evidences revealing important similarities between *S.* Napoli and typhoidal serovars (Huedo et al., 2017).

Objectives

To clarify the epidemiology of *S*. Napoli, which differs substantially from the great majority of the other serovars and for which the environment is considered the main reservoir. The data collected will be valuable to identify putative vehicles of human infection in order to assess possible control measures.

Project description

157 *S*. Napoli isolates from three partners (IZSVe, BfR and DTU) were collected and all *S*. Napoli sequence available on Enterobase (http://enterobase.warwick.ac.uk, 09/01/2018) were retrieved.

Methods and Results

Bioinformatic analysis included de-novo assembly (using SPAdes), automatic annotation (using Prokka) and core genome tree building (using Roary and Dendroscope for visualization). Samples were also characterized in terms of MLST, plasmid replicons, Salmonella Pathogenicity Islands (SPI), antimicrobial resistance genes (ARG), biocides and metal resistance genes (by BLAST), looking for features characterizing the clusters identified via phylogeny construction.

Preliminary results confirmed high genetic variability of this serovar (only 3477 out of 10499 genes were assigned to core genome), although only a few samples showed plasmid replicons and/or acquired ARGs. Samples tended to cluster according primarily to ST-type and biocide and metal resistance genes profile. All STs were spread among isolation sources and years of isolation, highlighting the challenge this serovar poses to trace its epidemiology and evolution.

Additional notes

The conclusion of the analyses is planned beyond the end of the ENAGE project period.

The project would be a valuable opportunity to collect important epidemiological evidences to clarify the epidemiology of an emergent serovar which is characterized by uncommon behaviour.

References

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control), 2017. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. EFSA Journal 2017;15(12):5077, 228 pp.

Fisher IS, Jourdan-Da Silva N, Hächler H, Weill FX, Schmid H, Danan C, Kérouanton A, Lane CR, Dionisi AM and Luzzi I, 2009. Human infections due to *Salmonella* Napoli: a multicountry, emerging enigma recognized by the Enter-net international surveillance network. Foodborne Pathogens and Disease, 6(5), 613-619. doi: 10.1089/fpd.2008.0206

Graziani C, Busani L, Dionisi AM, Caprioli A, Ivarsson S, Hedenström I and Luzzi I, 2011. Virulotyping of *Salmonella enterica* serovar Napoli strains isolated in Italy from human and nonhuman sources. Foodborne Pathogens and Disease, 8(9), 997-1003. doi: 10.1089/fpd.2010.0833

Graziani C, Luzzi I, Owczarek S, Dionisi AM and Busani L, 2015. *Salmonella enterica* serovar Napoli infection in Italy from 2000 to 2013: spatial and spatio-temporal analysis of cases distribution and the effect of human and animal density on the risk of infection. PLoS One, 10(11), e0142419. doi: 10.1371/journal.pone.0142419.

Huedo P, Gori M, Zolin A, Amato E, Ciceri G, Bossi A and Pontello M, 2017. *Salmonella enterica* Serotype Napoli is the First Cause of Invasive Nontyphoidal Salmonellosis in Lombardy, Italy (2010-2014), and Belongs to Typhi Subclade. Foodborne Pathogens and Disease, 14(3):148-151. doi: 10.1089/fpd.2016.2206

Sabbatucci M, Dionisi AM, Pezzotti P, Lucarelli C, Barco L, Mancin M and Luzzi I, 2018. Molecular and epidemiological analysis of remergent *Salmonella enterica* serovar Napoli, Italy, 2011-2015. Emerging Infectious Diseases, 24(3):562-565. doi: 10.3201/eid2403.171178

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA and Pevzner PA, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455-477. doi: 10.1089/cmb.2012.0021

Seemann T, 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069. doi: 10.1093/bioinformatics/btu153

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA and Parkhill J, 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), 3691-3693. doi: 10.1093/bioinformatics/btv421

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M and Rupp R, 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics, 8(1), 460.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ, 1990. Basic local alignment search tool. Journal of Molecular Biology, 215:403-410.

Appendix K.9

Project title	The ecological success of the monophasic variant of S. Typhimurium: a comparative genomic study
Consortium partner	IZSVe
Scientific lead	Antonia Ricci
Other people involved	Eleonora Mastrorilli, Sara Petrin, Alessandra Longo, Carmen Losasso, Lisa Barco

Purpose

The results of the study will contribute to clarify the success and diffusion of the *S.* 4,[5],12:i:- serovar and might provide valuable implications in terms of public health because of the recent emergence of this serovar.

Background

Over the past decades monophasic variant of *Salmonella* Typhimurium, *S.* 4,[5],12:i:- has been recognized as an emergent serovar worldwide for its rapid spread especially along the swine food chain. Since 2011 it has become the first serovar isolated both from human and veterinary sources in Italy and there has been an increasing interest toward the identification of the putative markers which could explain its epidemiological success. This was one of the main field of research of IZSVe during the last years and when ENGAGE started IZSVe had collected a panel of *S.* 4,[5],12:i:- strains to investigate through WGS to explore the epidemiological success of this relevant serovar.

Objectives

Although many studies have documented the ecological success of this serovar, few investigations have been conducted to explain this phenomenon from a genetic perspective. The aim of the study has been to identify the combination of different factors that could explain the eco-physiological role of this emergent serovar.

Project description

A comparative whole-genome analysis of 50 epidemiologically unrelated *S.* 4,[5],12:i:- isolates was performed. The isolates selected for the investigation were obtained from different sources (mainly cattle, pigs and poultry) over the last years (2010-2016) and characterized in terms of genetic elements potentially conferring resistance, tolerance and persistence characteristics.

Methods and Results

WGS was performed as described in Mastrorilli et al., 2018 (see below). The main genetic trait shared by the investigated strains was represented by heavy metals tolerance gene cassettes: most of the strains possess genes expected to confer resistance to copper and silver, whereas about half of the

isolates also contain the mercury tolerance gene *merA*. Functional studies were also performed to assess S. 4,[5],12:i:- capability to tolerate copper in the environment, in order to ascertain that the acquisition of heavy metal tolerance genes is useful for preventing the toxic effects of metals, thereby highlighting that this is a potential factor contributing to the success of this *Salmonella* serovar in farming environments.

In addition, the analysis of the distribution of type II toxin-antitoxin families indicated that these elements are abundant in this serovar, suggesting that this is another factor that might favor its successful spread. Phylogenetic analyses indicated a distinction among the investigated isolates based on the above described genetic traits, suggesting the involvement of different polymorphisms that give rise to multiple independent clones of *S.* 4,5,12:i:-.

Additional notes

This project was entirely conducted by IZSVe and ENGAGE provided an important technical and financial support to finalize this research.

All the activities planned were concluded. The results of the project have been described on a paper entitled Mastrorilli E, Pietrucci D, Barco L, Ammendola S, Petrin S, Longo A, Mantovani C, Battistoni A, Ricci A, Desideri A and Losasso C (2018) A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic *Salmonella* Serovar 4,[5],12:i:-. Front. Microbiol. 9:715. doi: 10.3389/fmicb.2018.00715

Appendix K.10

Project title	WGS of rare and unrecognised *Salmonella* serovars
Consortium partners	NVRI
Scientific lead	Magdalena Zając
Other people involved	Dariusz Wasyl

Purpose

1) To characterise *Salmonella* isolates causing difficulties in serotyping: incomplete or ambiguous antigenic structure and rare serovars (i.e. antigenic formulas observed occasionally in routine laboratory diagnostics) and from uncommon isolation sources (i.e. sheep, hedgehogs). 2) To set up an *in-house* database of reference sequences of rare and uncommon *Salmonella* serovars.

Background

Traditional *Salmonella* serotyping, classical tool for epidemiological study, does not always give unequivocal results. Additionally, it is time consuming method and needs laboratory experience. There are many isolates with incomplete antigenic structure i.e. lacking somatic antigens, one or both flagellar antigens that makes impossible to identify *Salmonella* serovar. Those isolates are often isolated along the food production chain and remain unrecognised, although they belong to common serovars of public health relevance. Our previous studies have shown that based on PFGE profiling or microarray analysis on-typable and autoagglutinating *Salmonella* strains might be clustered with S. Enteritidis, S. Hadar or S. Infantis (Hoszowski et al., Medycyna Wet 2011, 67:194-197; Zając at al., Acta Vet Hung 61:425-431). Moreover, some of isolates show atypical biochemical features which cause difficulties with classification to proper *Salmonella* subspecies or have the same antigenic structure but different biochemical properties (i.e. *S.* Paratyphi B *sensu stricto* and its d-tartrate positive var. Java). The other group of problematic isolates are rare serotypes often found in exotic pets like reptiles or hedgehogs. Due to the occurrence of rare or entirely new antigenic structure these strains can also be difficult to recognize properly. Limited number of available reference strains needed for QC of diagnostic sera broaden the uncertainty of identification of field isolates.

The next generation sequencing technology allows to look at isolates holistically and find information about the species, serovar, and subtype of bacteria in just one test.

Objectives

The project focused on the identification of problematic isolates where full antigen structure is unknown or small differences between serotypes do not allow to identify an isolate to *Salmonella* serovar and characterisation of rare serovars originated from exotic or uncommon sources (i.e. pet reptiles, sheep, hedgehogs). The research focused on developing of *Salmonella* identification methods and all aspects related with *Salmonella* characterisation. The project allowed to create a database of genomes which can be used as references and contribute to develop tools for rapid and reliable *Salmonella* identification.

Project description

WGS was performed on the isolates complying to one of the four criteria:

1. Serovar not identified due to missing antigens (O, H1 and/or H2) i.e.: 1,4,5:-:-; 4:d:-; 6,7:-:1,5; 6,7:c:-; 6,8:-:-; 9:-:-; 9:l,v:-; (monophasic *S*. Typhimurium excluded),

2. Serovar with ambiguous antigenic structure: *S*. Senftenberg (1,3,19:g,[s],t:-) and ***S*. Dessau** (1,3,15,19:g,s,t:-); *S*. Newport (6,8:e,h:1,2) and ***S.* Bardo** (8:e,h:1,2) (with particular emphasis on bolded serovars)

3. Rare serovars belonging to species and subspecies other than *Salmonella enterica* subsp*. enterica* and/ or deriving from uncommon sources like reptiles (native and exotic), exotic pets, and sheep,

4. Unknown antigenic structure and suspected of being new *Salmonella* serovar.

Methods and results

A total of 113 *Salmonella* strains were sequenced (Illumina MiSeq, HiSeq). WGS raw reads were processed with bbmerge v36.62, Trimmomatic v0.36, and SPAdes 3.9.0. Serovar presumptions were done with SeqSero 1.2, SISTR v1.0.1, *Salmonella*TypeFinder 1.4. CGE tool MLST 1.8 was used for sequence type identification. In some occasions phylogenetic analysis based on the concatenated alignment of SNPs was performed to compare strains within specific serovars.

1. WGS of 37 strains with incomplete antigenic structure make it possible to classify them to the most common and *Salmonella* control programme relevant serovars i.e. *Salmonella* 1,4,5:-:- recognised as *Salmonella* Typhimurium, *Salmonella* 9:-:- (*S.* Enteritidis), *Salmonella* 6,7:-:1,5, (*S*. Choleraesuis, *S*. Thompson), *Salmonella 6,7:z10:-* (*S*. Mbandaka), *Salmonella* 35:-:- (*S*. Monschaui). Sequence types of the strains allowed their allocation in the common clones i.e. *S.* Enteritidis ST 11, and *S*. Mbandaka ST 413 occurring in food chain in Poland.

2. Twenty-six stains were sequenced. It has been confirmed that *Salmonella* Bardo should be actually included into *Salmonella* Newport. The isolates occurring occasionally in food chain in Poland are considerably diverse (SNP tree) and represent several sequence types (ST 31, ST 118, ST 166). Similar conclusions were drawn for *Salmonella* Dessau that should be included into *S.* Senftenberg (ST14, ST 210).

3. Relevant gene identification confirmed serovar identification of 50 isolates of strange or uncommon origin i.e. European grass snake (*Salmonella* IIIb 28:z10:z, unknown ST; IIIb 38:r:z:[z57], ST 645; *Salmonella* Sunnycove, unknown ST), pet geko (*Salmonella* II 16:m,t:[z42], unknown ST). Noteworthy, few available sheep isolates were identified as *Salmonella* IIIb (diarizonae) 61:k:1,5 (ST 432) – the serovar being considered sheep specific pathogen. As far as we are concerned this is the first report of *Salmonella* IIIb 61:k:1,5 (ST 432) in Poland.

4. None of the sequenced sample revealed previously unknown antigenic formula.

References

Hoszowski A, Lalak A, Zając M, Samcik I, Skarżyńska M, Wnuk D, Wasyl D. 2011 [Relationship of rough *Salmonella* strains with representatives of some serovars found in animals]. Medycyna Wet 67:194-197.

Zając M, Hoszowski A, Wasyl D. 2013. Identification of common, non-typable and autoagglutinating *Salmonella* strains with Premi®Test Salmonella Assay. Acta Vet Hung 61:425-431.

Additional notes

Conclusions were implemented in routine diagnostic activity of NRL *Salmonella*

Zając M., Bomba A., Skarżyńska M., Giza A., Wasyl D. Whole genome sequencing of "non-existing" *Salmonella* serovars, International Symposium Salmonella and Salmonellosis, Saint Malo, France, 24-26 Sept. 2018, submitted

Appendix K.11

Project title	Characterization of *mcr*-1 positive *E. coli* isolated from food-producing animals in Poland
Consortium partners	NVRI, DTU
Scientific lead	Magdalena Zając
Other people involved	Dariusz Wasyl, Valeria Bortolaia, Pimlapas Leekitcharoenphon, Rene Hendriksen, Paweł Sztromwasser

Purpose

To assess prevalence and possible dissemination of *mcr*-1 positive *E. coli* among animals in Poland

Background

The emergence of transmissible plasmid-mediated colistin resistance (*mcr*-1) has posed a threat to effective use of polymyxins, which are considered the last line of defence against multidrug and carbapenem resistant Gram-negative bacteria. Little is known about the prevalence of colistin resistance and the *mcr* gene in livestock in Poland.

Objectives

The objective of this study was to characterize mcr-1-harbouring *E. coli* isolated from animals in Poland and determine the possible dissemination pathways of detected colistin-resistance mechanisms in tested isolates.

Project description

Analysis of MICs for colistin allowed for selection of 128 suspected colistin-resistant isolates. The presence of the *mcr*-1 gene was confirmed (PCR evaluation for isolates meeting the selection criteria of MIC≥2mg/L for *E. coli* isolated in 2014-2016 and MIC>2mg/L for isolates identified in 2011-2013) in 80 out of suspected strains. They were recovered from turkeys (n=64), broilers (n=12), laying hens (n=2), cattle (n=1) and pig (n=1).

Methods and results

mcr-1 positive strains were sequenced (Illumina MiSeq, HiSeq). The raw reads were processed using bbmerge v36.62 to merge overlapping reads and Trimmomatic v0.36 to trim adapters and low quality reads. Merged reads and trimmed not-merged pairs were used to generate assembly contigs using SPAdes 3.9.0. Sequences were analyzed for their content of resistance genes, plasmid replicons and virulence genes by using the CGE Web tools (https://cge.cbs.dtu.dk/services/) ResFinder 3.0, VirulenceFinder 1.5, PlasmidFinder 1.3 and pMLST v1.4 for typing of IncHI2 plasmids. MLST profiling was performed using MLST 1.8. Phylogenetic tree was constructed by complete linkage clustering using sequence-similarity distance matrix. The distance matrix was generated by global pairwise

MUMmer 3.23 alignments between samples' contigs (automated by CONOCOCT 0.4.0 script dna_diff_distmatrix.py). I-Tol web-based tool was used to display the tree. The following analyses were performed:

- generation of phylogenetic tree of *mcr*-1 positive *E. coli*;
- determination of MLST of isolates;
- identification of antimicrobial resistance genes;
- detection of plasmids;
- defining of virulence genes;
- assessing possible location of the *mcr-1* gene.

All tested strains carried *mcr*-1 gene and no other *mcr* genes were found. The results showed an increase of colistin resistance with time and a sudden increase of *E. coli* harbouring *mcr*-1 in turkey flocks in 2016. Whole-genome sequencing showed high diversity of ST types, various plasmids (i.e. *Inc*Q1, *Inc*X4, *Inc*HI2, *Inc*FII, *Inc*I1, *Inc*FIB) and a complex resistance background. ESBL genes (i.e. $bla_{CTX\text{-}M\text{-}1}$, $bla_{CMY\text{-}2}$, $bla_{SHV\text{-}12}$, $bla_{TEM\text{-}52}$) and quinolone resistance genes (*qnr*S1, *qnr*B19) were present in a wide variety of *E. coli* STs. Chromosomal point mutations for fluoroquinolone resistance in *gyrA* (S83L, D87N) and *parC* (S80I) were the most frequently identified. All the above mentioned cephalosporin and quinolone resistance genes were previously detected in *E. coli* isolated from animals in Poland and tested with non-NGS methods. The *mcr*-1 genes were mostly found on the same contig as the *Inc*X4 or *inc*HI2 replicons, and mutations in *pmr*B were found along with the *mcr*-1 gene. One isolate carried *mcr*-1 located on the chromosomal sequence and thus suggesting possible vertical spreading of the resistance. Current study is the first confirmed case of presence of the *mcr*-1 gene in bacteria of animal origin in Poland. The simultaneous prevalence of *mcr*-1 and other genes conferring resistance to fluoroquinolones and cephalosporins in healthy food-producing animals draws attention to rational use of colistin in veterinary practice. As the abundance of resistant bacteria may be underestimated, rapid spread of *mcr*-1 among turkeys could forecast possible appearance of the gene in food chain in Poland.

References

Zając M, Sztromwasser P, Wasyl D, Hoszowski A. 2017. Whole genome sequencing characterisation of *mcr*-1 positive *Escherichia coli* isolated from turkeys and chickens, Proceedings of 2nd International Caparica Conference in Antibiotic Resistance, (Caparica, Portugal, 12th – 15th June, 2017), p 188.

Zając M, Sztromwasser P, Bortolaia V, Leekitcharoenphon P, Cavaco L, Ziętek-Barszcz A, Hendriksen R, Wasyl D, Prevalence and characterization of *mcr*-1 positive *E.coli* isolated from food-producing animals in Poland (manuscript under preparation).

Additional notes

-

Appendix L – The ENGAGE Proficiency Test Report 2016

THE ENGAGE PROFICIENCY TEST REPORT 2016

Oksana Lukjancenko, Susanne Karlsmose Pedersen, Rene S. Hendriksen

1. edition, March 2017.

Technical University of Denmark, National Food Institute, Research Group of Genomic Epidemiology, Kgs. Lyngby, Denmark

Contents

1. Introduction

The main objective of this proficiency test (PT) is to facilitate the production of reliable laboratory results of consistently good quality within the area of whole genome sequencing (WGS).

The PT evaluates the consistency and robustness of ENGAGE consortium members' ability to perform deoxyribonucleic acid (DNA) extraction, library preparation, the WGS, and assembly following different laboratory protocols, software tools, and sequence platforms for the reliability of submitted sequence data to the public repositories. This ensures harmonization and standardization in WGS and data analysis, with the aim to produce comparable data for the ENGAGE initiative. To meet these objectives, the laboratory work and analyses performed for this PT should be performed using the methods routinely employed in the individual laboratories.

The PT consists of a "Wet-lab" component targeting three common bacterial pathogens. The Wet-lab components assess the laboratories ability to perform DNA preparation, sequencing procedures and, if laboratories routinely do so, the analysis of epidemiological markers; Multi Locus Sequence Typing (MLST) and antimicrobial resistance (AMR) genes.

The individual laboratory data are confidential and only known by the participating laboratory and the PT organizers (DTU Food).

Materials and Methods

Participants

A pre-notification to announce the ENGAGE proficiency test was distributed on the 12th July 2016 by e-mail to the eight ENGAGE consortium partners. Seven of the eight partners signed up and participated in the PT. Only, the National Institute of Public Health – National Institute of Hygiene in Poland did not participate as they have not initiated in-house WGS. Some of the seven partners however, only took part in testing a subset of the target organisms after agreement with the PT organizers.

Strains

Two strains of *Campylobacter jejuni*, *Listeria monocytogenes*, and *Klebsiella pneumonia* were selected for the wet-lab in 2016. In a GMI end-user analysis of what species to target, *Campylobacter* and *Listeria* have been indicated being of interest (Moran-Gilad et al., 2015). *Campylobacter* was selected for this PT due to its many repeats and rearrangements and *Listeria* due to it being part of many genomic pilot projects and it's genetically heterogeneous with limited repeats and rearrangement. One of the *Listeria* strains belonged to a less virulent MLST – ST-121, whereas the other strain was of to a known virulent type, ST-2. We also included *Klebsiella* due to its many resistance genes for evaluating if the detection of these as can be used to indicate the quality level of the sequencing.

Individual sets of the strains were lyophilized as KWIK STIKs by Microbiologics, St. Cloud, Minnesota, USA and the corresponding DNA were purified and pooled by DTU-Food prior to distribution in individual vials for each participant.

To better be able to assess the differences in the sequences generated by the participants, each of the six strains in the Wet-lab component were sequenced on the PacBio to get a closed reference genome. This was done by creating 10kb template libraries using "10kb DNA Template Prep Kit 1.0" from Pacific Biosciences, which were then sequenced using C2 chemistry on single-molecule real-time (SMRT) cells with a 180min collection protocol. The data was then de novo assembled using the Hierarchical Genome Assembly Process (HGAP) within the Pacific Biosciences SMRTAnalysis software package. Polishing and finishing the genome were performed with custom python scripts, Quiver and Gepard, a dot plot tool to identify overlapping regions.

Distribution

On 24th October 2016, bacterial strains in agar stab cultures together with the corresponding purified and dried DNA and a welcome letter were dispatched in double pack containers (class UN 6.2) to the participating laboratories according to the International Air Transport Association (IATA) regulations as UN3373, biological substances Category B.

Procedure

The protocol was made available on the website (http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016) allowing the PT participants access to all necessary information at any time. Additional relevant information was distributed by email directly to the participants.

The protocol presented instructions as to the handling of the received bacterial cultures and DNA.

Participants were requested to capture information in relation to the questions presented in the SurveyMonkey.

Deadline for submission of results was initially set for 14th December 2016 but was extended to 13th January 2017.After this date, participants, #93 and #104 who had not yet submitted results according to the level of their sign-up, were approached to confirm if they were planning on submitting results. By the beginning of February 2017, all relevant data was captured and the data analysis was instigated. This report summarizes the results and allows for ensures full anonymity for the participants, as only the PT-organizers has access to the individual results.

2.3.1. SurveyMonkey

Apart from three questions relating to the contact information of the participant, 40 questions were asked focused on the storage of bacterial cultures and DNA prior to analysis, the cultivation and DNA extraction procedure, the quality assurance parameters applied, details related to the sequencing and analysis of the obtained sequencing data.

2.3.2. Sequencing

The participants uploaded raw sequence files in fastq format. The reads were *de novo* assembled applying the standard assembly pipeline used by the web-services from Center for Genomic Epidemiology (CGE) https://cge.cbs.dtu.dk//services/all.php, except for the reads which were not trimmed prior to the assembly.
For the raw reads, the following QC metrics were calculated:

- Number of reads that map to reference chromosome
- Proportion of reads that map to reference chromosome out of all reads that map to total reference DNA
- Coverage, total reference DNA. The number of reads mapping to the total reference DNA multiplied with the average length of the reads divided by the total size of the reference genome

For the assemblies, the following QC parameters were calculated:

- Size of assembled genome
- Size of assembled genome per total size of DNA sequence
- Total number of contigs
- N50 (defined as the length of the shortest contig, in the set of largest contigs that represents at least 50% of the assembly)

In addition to the calculation of the above QC metrics and parameters, participants were requested to provide the identification of the strains corresponding MLST and AMR genes to support the assessment of the sequence quality. Participates identified the MLSTs and AMR genes using the software of their choice. To assess the proficiency of the participants, the PT organizers used a command line version of the CGE MLST-Finder v.1.7 Larsen et al., 2012) and ResFinder 2.1 (Zankari et al., 2012) (Threshold for %ID = 98% and HSP/Query length = 60%) including the CGE standard assembly pipeline on the participant's raw reads to compare the results with those reported by the participants.

Results

Participation

Seven laboratories responded to the pre-notification and were enrolled in the ENGAGE PT. When the deadline for submitting results was reached, all seven laboratories had uploaded data. Seven partners, #104, #114, #115, #77, #82, #93, and #95 submitted raw reads of both the culture and the DNA for both *Campylobacter* strains. Only four partners, #104, #115, #77, and #82 submitted raw reads of both the culture and the DNA for both *Listeria* strains and five partners, #104, #114, #77, #82, and #93 submitted raw reads of the *Klebsiella* cultures and the DNA.

Method description

The bacterial cultures were stored at 4°C by 86% (n = 6) of the participants prior to the analysis. In addition, one participant, #104 (14%) stored the reference material at -20°C.

Four participants (57%) stored the DNA in the time between reception and processing at room temperature (5 days by #115, 12 days by #77, 14 days by #82 and 41 days by #114) whereas the remaining three participants, #93, #95 and #104 stored the DNA at 4°C.

All seven participants inoculated the bacterial cultures onto various types of blood agar. The *Listeria* and *Klebsiella* strains were incubated at 37°C between 16 to 24 hours in contrast to *Campylobacter* which were incubated at 42°C for 48 hours.

By five partners, the Genomic DNA was extracted from the Gram negative and positive using a number of different commercially available kits including, Easy-DNA and PureLink Genomic DNA Mini Kit (Gram negative) from Invitrogen, Minikit (Gram negative) and QIAamp DNA Mini kit from Qiagen, Charge Switch gDNA Mini Bacteria Kit (Gram positive) and Genomic Mini from A & A Biotechnology. Two of the participants have modified the used Gram positive protocols by lysostaphin treatment prior to extraction. Two partners used a commercially available automatically DNA purification instrument/ robot, the MagNA Pure LC / MagNA Pure LC DNA Isolation Kit III (Bacteria, Fungi) from Roche and the QIAsymphony/ DSP DNA Mini Kit from Qiagen.

DNA concentrations (ng/µl) of the bacterial cultures and DNA were determined prior to library preparation on a Qubit by four partners. In addition, one participant used the Nanodrop and another participant the GloMax® 96 Microplate Luminometer (QIAsymphony) and a third a quantifluor kit read on POLARstar Omega plate reader.

For the *Campylobacter* cultures, the DNA concentration ranged from 0.26 to 80 ng/µl and from 0.22 to 647.39 ng/µl for the provided DNA (Table L.1). For the *Listeria* culture, the DNA concentration ranged from 0.28 to 92.05 ng/µl and from 0.24 to 33.52 ng/µl for the DNA. The DNA concentration ranged from 0.18 to 34.3 ng/µl and from 0.24 to 58.3 ng/µl for the *Klebsiella* bacterial culture and DNA, respectively (Table L.1).

For the *Campylobacter* culture, the total amount of DNA ranged from 0.001 to 4.8 µg and from 0.001 to 3.12 µg for the provided DNA (Table L.2). For the *Listeria* culture the total amount of DNA ranged from 0.001 to 4.88 µg and from 0.001 to 3.11 µg for the total amount of DNA. The total amount of

DNA ranged from 0.001 to 1.43 µg and from 0.001 to 3.43 µg for the *Klebsiella* culture and DNA, respectively (Table L.2). Laboratory #77, consistently reported the concentrations of 0.001.

All seven participants responded to the method applied to measure the DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation for bacterial cultures and DNA received. For bacterial cultures, two (29%) of the laboratories used the Nanodrop, one used the Qubit, another one the BioPhotometer plus (Eppendorf), and a third one quantifluor kit read on POLARstar Omega plate reader. In addition, two (29%) did not measure the DNA quality. For the DNA received, the laboratories used the same method to measure the DNA concentration except for one of the participants that reported not measuring the DNA of the cultures, and which used the Nanodrop.

Up to five of the laboratories depending on participation reported the measurement of the DNA quality (e.g. RIN or 260/280 ratio) for bacterial cultures and DNA received (Table L.3). Among the five laboratories providing data of the DNA quality for the cultures, the level ranged from about 1.47 to 12.1 (Table L.3).

Four participants reported the measurement of the DNA quality (260/230 ratio) for bacterial cultures and DNA received (Table L.4). For the cultures and received DNA, the DNA quality ranged from 0.2 to 2.4 (Table L.4).

Two out of the seven laboratories assessed the quality visually on an agarose gel.

Of the seven participants, five used the Illumina Nextera XT DNA sample preparation kit FC-131-1024 (n = 2) and FC-131-1096 (n = 2) and one indicated using the Illumina NEB Next Ultra DNA Library prep kit E6040L for the preparation of the sample library before sequencing. Two participants using the Illumina Nextera XT DNA kit FC-131-1024 or FC-131-1096, respectively indicated using this in combination with the Nextera XT Index kit FC-131-1001 or FC-121-1012. In addition, one participant did not indicate the cat no. but the lot no.

The genomic DNA was prepared for pair-end sequencing by all seven (100%) participants. The libraries were sequenced by five participants (71%) using an Illumina MiSeq platform whereas two used the HiSeq 2000 or the HiSeq 2500 platforms, respectively. The read length of the sequences was set between 100 (n = 1), 250 (n = 1), 251 (n = 3) and up to 300 bp (n = 1). The reads were trimmed before upload by one, #115 out of the seven participants using trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Five participants indicated that if assembled by themselves, three would have used SPAdes http://bioinf.spbau.ru/spades, one would have used the PATRIC provided tool (https://www.patricbrc.org) ; Assembly Strategy: FullSpades, output file "contigs.fa" and finally one have used Assembler 1.2: https://cge.cbs.dtu.dk/servicesAssembler available from CGE.

Sequencing, MLST, and antimicrobial resistance genes

For *Campylobacter* GMI16-001 the expected MLST was ST7426 which was found by all laboratories except for Laboratory #114 that had mixed up the two strains GMI16-001-BACT and GMI16-002-BACT as well as GMI16-001-DNA and GMI16-002-DNA, explaining the incorrect MLSTs. Two laboratories, #95 and #114 did not report MLST data (own tool) but these were provided by PT-organizer (CGE tool) and found correct (Table L.5).

The *Campylobacter*; GMI16-001 was pan-susceptible why no resistance genes were expected. Laboratory #114 reported however, resistance data matching the profile of GMI16-002 due to the mix up (Table L.6).

The MLST ST6238 was expected in *Campylobacter* strain; GMI16-002. This was reported by all participants except for Laboratory #114 due to the above reported mistakes (Table L.5).

A very high degree of concordance was observed between the reported resistance genes detected by own tools and the CGE reference tool and between culture and DNA samples. Only four participants reported what own tool being used to identify the resistance genes, #77, #82, and #93 used the CGE

ResFinder whereas #104 used Blastn. Some of the resistance genes, were determined "like" which indicate that the homology to the reference genes were less than 100% which is often seen due to minute sequencing errors. The gene *aph*(2'')-like was reported by a number of laboratories. In contrast, the CGE tool did not detect this specific gene which doesn't mean that it is not present. It merely indicate that the commandline version of the CGE ResFinder tool did not pick up this gene most likely due to a higher threshold in homology than used by the laboratories. Laboratory #82 reported chromosomal point mutations which are not yet included the commandline version of the CGE ResFinder tool why this very well could be true. Running the commandline version of the CGE ResFinder tool for the genome of *Campylobacter* strain; GMI16-002 submitted by laboratory #114 showed resistance genes that do not match any of the expected profiles of the PT strains (Table L.6).

Only four laboratories, #77, #82, #104, and #115 tested the two *Listeria* strains; GMI16-003 and GMI16-004. In all cases, the four laboratories managed to identify the correct and expected MLST ST-2 and ST-121, respectively (Table L.5).

The two *Listeria* strains were both pan-susceptible and no resistance genes were reported nor identified using the commandline version of the CGE ResFinder tool.

Five laboratories, #77, #82, #93, #104, and #114 tested the *Klebsiella* strains, GMI16-005. The commandline version of the CGE MLSTFinder tool was used to test the submitted genome, GMI16-005-DNA laboratory #114 which didn't submit own data. In all cases, the laboratories managed to identify the correct and expected MLST ST-512 (Table L.5).

The same laboratories were involved in testing the *Klebsiella* strains, GMI16-006. For this strain laboratories #114 didn't submit own data. All MLST profiles were correct, ST-15 (Table L.5).

Both of the *Klebsiella* strains were multidrug resistant harbouring a number of resistance genes (Tables L.7-L.8). *Klebsiella* strains, GMI16-005 were found to contain the following genes, *aad*A2, *aac*(6')-Ib), $bla_{\text{TEM-1A}}$, $bla_{\text{KPC-3}}$, $bla_{\text{OXA-9}}$, $bla_{\text{SHV-11}}$, *oqx*A, *oqx*B, *aac*(6')Ib-cr, *fos*A, *mph*(A), *cat*A1, *sul*1, and *dfr*A12. Most of the genes were identified by both own and CGE tools indicated by a very high concordance. Several of the laboratories report the genes being with a lower homology than the reference gene indicated by being determined "like". The mutation, *aac*(6')Ib-cr was not identified by laboratory #82 using own tools for both the culture and DNA in contrast to the CGE tool. The laboratory however, identified the presence of the gene, *aac*(6')Ib as all did. This indicate that the laboratory might have used another tool not able to identify this mutation in the *aac*(6')Ib gene. Similarly, the CGE tool was not able to detect neither the *aac*(6')Ib nor the mutation *aac*(6')Ib-cr in GMI16-005-DNA for laboratory #104 indicating a potentially truncated gene. Almost all of the laboratories identified the gene, *fos*A in a "like" version. The commandline version of the CGE ResFinder tool did not pick up this gene most likely due to a higher threshold in homology than used by the laboratories (Table L.7).

The *Klebsiella* strains, GMI16-006 contained the following genes, *aad*A1, *aac*(6')-Ib, *aac*(3)-Iid, *aph*(3')-Via, *str*A, *str*B, $bla_{\text{NDM-1}}$, $bla_{\text{OXA-9}}$, $bla_{\text{CTX-M-15}}$, $bla_{\text{SHV-1}}$/ $bla_{\text{SHV-28}}$, $bla_{\text{TEM-1b}}$/$bla_{\text{TEM-1a}}$, *qnr*S1, *oqx*B, *oqx*A, *aac*(6')Ib-cr, *sul*2, *tet*(D), *dfr*A14, and *fos*A. The concordance was very high between the laboratories testing the strain GMI16-006. In two incidences, the commandline version of the CGE ResFinder tool identified $bla_{\text{TEM-1a}}$ whereas all "own" testing as well as the remaining testing by the CGE tool identified the gene, $bla_{\text{TEM-1b.}}$ The difference between the two genes is only a few SNPs why the error is often observed. Inconsistences in detection of the gene bla_{SHV} gene were observed. Some laboratories couldn't distinguish the type of bla_{SHV} and reported $bla_{\text{SHV-28}}$ or $bla_{\text{SHV-28}}$, respectively. Consistency however, between "own" and CGE data was seen. The same explanation given for the bla_{TEM} gene also accounts the bla_{SHV} gene. Almost all of the laboratories identified the gene, *fos*A in a "like" version. The commandline version of the CGE ResFinder tool did not pick up this gene most likely due to a higher threshold in homology than used by the laboratories. Laboratory #93 reported the detection of the $bla_{\text{LEN-12-Like}}$ gene not reported by others (Table L.8).

Sequencing, Quality markers

All seven laboratories submitted sequencing data for the *Campylobacter* GMI16-001 and GMI16-002 related to the quality metrics and parameters from both the received bacterial culture and corresponding DNA. For Listeria GMI16-003 and GMI16-004, four laboratories participated, #77, #82, #104, and #115 both the received bacterial culture and corresponding DNA except for laboratory #77 which didn't submit data for the corresponding DNA of GMI16-004. In testing the *Klebsiella* strains, GMI16-005 and GMI16-006, the following laboratories participated #77, #82, #93, #104, and #114 submitting data for both the bacterial culture and corresponding.

The quality metrics and parameters of GMI16-001-BACT, GMI16-002-BACT, GMI16-001-DNA, and GMI16-002-DNA from laboratory #114 were excluded the analysis due to mixing up the two strains.

Initially, the quality markers were evaluated for potential contamination which revealed that all genomes were of only one species.

The medians of the number of reads mapped to the reference DNA sequences was somehow consistent between the three species with a tendency of having higher medians for the DNA than the BACT samples (Figure L.1). The 25% upper and lower quartiles ranged largely for the *Listeria* and *Klebsiella* species compared to *Campylobacter*. In general, two laboratories, #93 and #115 were determined outliers with a high number of reads mapped to the reference DNA sequences for all species compared to the other laboratories (Figure L.1). This can be explained by the used sequencing platform e.g. Hiseq 2000 and Hiseq 2500 in contrast to Miseq being used by all other laboratories. In addition, these platforms often provide more reads and of shorter length e.g. 100bp as indicated by laboratory #93. The lowest observed values were of laboratory #104 and #114 (Figure L.1).

The proportion of reads produced which map directly to the closed genome of the same strain should not exceed more than 100% indicating an error e.g. contaminations. The medians of the proportion of reads produced which map directly to the closed genome were almost 100% for both the DNA and the culture of the two *Campylobacter* genomes. A very little range of the 25% upper and lower quartiles were observed. This indicated that all laboratories performed equally well with the exception of laboratory #115. For GMI16-001-BACT, laboratory #115 only had 56.2% of reads mapped to the reference DNA sequence (Figure L.2). The median of the proportion of reads produced which map directly to the closed genome of the *Listeria* GMI16-003-BACT were as well almost 100% with a tight upper and lower quartile centered around. In contrast, for the GMI16-003-DNA, the upper and lower quartile were much larger but still close to 100% indicating that all reads produced map to the reference. Lower proportions of reads produced which map directly to the closed genome were observed for *Listeria* GMI16-004 with a median of close to 89% for BACT and 86% for DNA. The reason for this might be due to a high number of plamids. The medians of the proportions of reads produced which map directly to the closed genome for *Klebsiella* GMI16-005 and GMI16-005 were about 95% or greater, especially for GMI16-005 which indicates a nice fit of the reads to the reference. In general, the proportions of reads produced which map directly to the closed genome were lower for the laboratories, #93, #104, and #115 than the others participating (Figure L.2).

The total number of contigs assembled should ideally be less than 1000 indicating good quality – the lower the better. For *Campylobacter* GMI16-001 and *Campylobacter* GMI16-002, the medians are between 100 and 150 contigs with a tight 25% quartile fit except for GMI16-001-BACT and GMI16-002-DNA where laboratory #104 produced 459 and 2.131 contigs, respectively and being considered an outlier. The 25% quartiles are much broader for *Listeria* GMI16-003 and GMI16-004 with an overall median less than 250 contigs indicating some unexpected difficulties sequencing *Listeria*. Similarly, the *Klebsiella* genomes of GMI16-005 and GMI16-006 revealed medians below the same number of 200 contigs. For the DNA samples, the 25% quartiles are really tight compared to the BACT, indicating that the problems can be related to the DNA purification step of the bacteria. This seems to be a general observation (Figure L.3).

The size of the assembled genomes was observed to match the expected size of the species with *Campylobacter* being around 2mb, *Listeria* about 3mb, and *Klebsiella* of around 5mb. For the *Campylobacter* genomes, laboratory #104 was considered an outlier with a size of 2.077.671bp (107.92%) for GMI16-001-BACT and 3.260.167bp (171.72%) for GMI16-002-DNA (Figures L.4-L.5). For *Listeria*, the size the assembled genomes as well as the proportion of the size to the reference DNA sequences were much broader with larger 25% quartiles especially for the DNA samples. In contrast, the proportion of the size to the reference DNA sequences of *Klebsiella* GMI16-005 were in average 99.3% with an outlier of 111.67% (laboratory #104) (Table L.4). Larger 25% quartiles were observed for *Klebsiella* GMI16-006 but still close to the expected size of the species genome and with an almost 100% in proportion to the reference DNA sequence (Tables L.3-L.4).

The N50 length is defined as the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs. A N50 more than 15000 normally indicate good quality and were obtained by all laboratories for all of the genomes. The lowest N50 value observed was 83.000 and by Laboratory #104 (Figure L.6).

The depth (bp) of the coverage is calculated based on the number of bps sequenced divided by the total size (both chromosome and plasmids) of the closed genome (same strain). This number can be rounded to the nearest integer. In essence this number describes the number of times the sequenced bps covers the reference DNA and is often ended with an "x" (e.g. 30x) which also serve as a good average number in depth. All of the laboratories for all genomes were observed to have an overall depth of between 50X to 100X which is ideal (Figure L.7).

Discussion

The majority of the submitted MLST data were correct and in line with the expected value. The results of MLST analysis revealed a systematic error for participant #114 when submitting the data causing a mix up of the test genomes for the MLST and resistance genes prediction. However, the MLST was correct for all PT strains when re-analysed using the CGE reference method.

Most of the submitted AMR genes were in concordance with the expected results. Some deviations however were observed mostly due to the tools, own or CGE reference lower threshold setting ignoring genes with a lower homology.

One of the objectives for the ENGAGE PT was to assess a range of quality markers to evaluate the performance by the consortium partners. Overall, the PT test show that all laboratories perform satisfactory with the exception of laboratory #104 which in general produced a low number of reads, a lower percentage of mapping reads to the references, a high number of contigs, a high size of the assembly, and a high proportion in the size of the assembly per reference sequence. A few other laboratories could benefit from an assessment of own sequencing quality including laboratory #114. It is noteworthy to mention that the lower quality of the sequeces produced by laboratory #104 did not affect the prediction of MLSTs nor resistance genes.

Laboratory #104 have indicated that they tried to select standard parameters of sequencing (routinely used) with a depth oscillates of about 30x which can affect the results of laboratory #104. In addition, the laboratory submitted trimmed sequences as indicated in the protocol but failed to remove adapters which normally are removed by the platform itself. This might have affected the quality of the sequences as the PT organizers didn't enhance any of the submitted data. The PT organizers offered the laboratory #104 to re-submit data with removed adapters but this was not possible due to the timeline and deliverable of this report.

Conclusions

The pilot PT was a useful exercise as it allowed ENGAGE consortium partners to assess the quality of own data as well as to identify critical points for improvement. In general, all data were satisfactory but the PT organizer encourage especially laboratory #104 to upload data which has removed adapters as well as ensuring the genomes being matched to the identical reference genome to avoid wrong prediction of the MLST and resistance genes.

References

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM and Lund O, 2012. Multilocus sequence typing of total-genome-sequenced bacteria. Journal of Clinical Microbiology, 50(4):1355-1361.

Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E and Hendriksen RS, 2015. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infectious Diseases, 15:174-0902.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM and Larsen MV, 2012. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy, 67:2640-2644.

THE ENGAGE PROFICIENCY TEST REPORT 2016

TABLES AND FIGURES

Table L.1: The low and high range of DNA concentration (ng/µl) measured for both the bacterial cultures and DNA received

	Low range (ng/µl)	High range (ng/µl)
GMI16-001-BACT *(Campylobacter)*	0.26	64.8
GMI16-002-BACT *(Campylobacter)*	0.27	80
GMI16-001-DNA *(Campylobacter)*	0.27	647.39
GMI16-002-DNA *(Campylobacter)*	0.22	90.89
GMI16-003-BACT *(Listeria)*	0.28	92.05
GMI16-004-BACT *(Listeria)*	0.28	68.7
GMI16-003-DNA *(Listeria)*	0.25	33.52
GMI16-004-DNA *(Listeria)*	0.24	23.77
GMI16-005-BACT *(Klebsiella)*	0.18	34.3
GMI16-006-BACT *(Klebsiella)*	2.2	28.6
GMI16-005-DNA *(Klebsiella)*	0.25	52.18
GMI16-006-DNA *(Klebsiella)*	0.24	58.3

Table L.2: The low and high range of total DNA amount (µg) measured for both the bacterial cultures and DNA received

	Low range (µg)	High range (µg)
GMI16-001-BACT *(Campylobacter)*	0.001*	3.88
GMI16-002-BACT *(Campylobacter)*	0.001*	4.8
GMI16-001-DNA *(Campylobacter)*	0.001*	2.28
GMI16-002-DNA *(Campylobacter)*	0.001*	3.12
GMI16-003-BACT *(Listeria)*	0.001*	4.52
GMI16-004-BACT *(Listeria)*	0.001*	4.88
GMI16-003-DNA *(Listeria)*	0.001*	3.11
GMI16-004-DNA *(Listeria)*	0.001*	1.79
GMI16-005-BACT *(Klebsiella)*	0.001*	1
GMI16-006-BACT *(Klebsiella)*	0.001*	1.43
GMI16-005-DNA *(Klebsiella)*	0.001*	3.43
GMI16-006-DNA *(Klebsiella)*	0.001*	2.45

* All values from #77.

Table L.3: The low and high range of the measured DNA quality (e.g. RIN or 260/280 ratio) for both the bacterial cultures and DNA received

	Low range	High range
GMI16-001-BACT *(Campylobacter)*	1.47	3.28
GMI16-002-BACT *(Campylobacter)*	1.84	2.54
GMI16-001-DNA *(Campylobacter)*	2	3.55
GMI16-002-DNA *(Campylobacter)*	1.75	2.26
GMI16-003-BACT *(Listeria)*	1.7	1.93
GMI16-004-BACT *(Listeria)*	1.77	1.91
GMI16-003-DNA *(Listeria)*	1.82	1.92
GMI16-004-DNA *(Listeria)*	1.78	1.86
GMI16-005-BACT *(Klebsiella)*	1.85	11.2
GMI16-006-BACT *(Klebsiella)*	1.73	10.2
GMI16-005-DNA *(Klebsiella)*	1.72	12.1
GMI16-006-DNA *(Klebsiella)*	1.78	13

Table L.4: The low and high range of the measured DNA quality (260/230 ratio) for both the bacterial cultures and DNA received

	Low range	High range
GMI16-001-BACT *(Campylobacter)*	0.79	2.4
GMI16-002-BACT *(Campylobacter)*	1.71	2.24
GMI16-001-DNA *(Campylobacter)*	0.54	2.3
GMI16-002-DNA *(Campylobacter)*	0.21	1.62
GMI16-003-BACT *(Listeria)*	1.31	1.73
GMI16-004-BACT *(Listeria)*	1.35	1.95
GMI16-003-DNA *(Listeria)*	0.32	1.75
GMI16-004-DNA *(Listeria)*	0.2	1.83
GMI16-005-BACT *(Klebsiella)*	1.36	1.92
GMI16-006-BACT *(Klebsiella)*	1.22	1.76
GMI16-005-DNA *(Klebsiella)*	0.24	1.76
GMI16-006-DNA *(Klebsiella)*	0.32	1.6

Table L.5: Determined MLST for both the bacterial culture and DNA received

			GMI16-001		GMI16-002		GMI16-003		GMI16-004		GMI16-005		GMI16-006	
	Participant		Expected MLST	Obtained MLST	Expected MLST	Obtained MLST	Expected MLST	Obtained MLST	Expected MLST	Obtained MLST	Expected MLST	Obtained MLST	Expected MLST	Obtained MLST
BACT	#77	Own tool	ST-7426	7426	ST-6238	6238	ST-2	2	ST-121	121	ST-512	512	ST-15	15
		CGE tool		7426		6238		2		121		512		15
	#82	Own tool		7426		6238		2		121		512		15
		CGE tool		7426		6238		2		121		512		15
	#93	Own tool				6238						512		15
		CGE tool		7426		6238						512		15
	#95	Own tool												
		CGE tool		7426		6238								
	#104	Own tool		7426		6238		2		121		512		15
		CGE tool		7426		6238		2		121		512		15
	#114*	Own tool												
		CGE tool		**6238**		**7426**						512		
	#115	Own tool		7426		6238		2		121				
		CGE tool		7426		6238		2		121				
DNA	#77	Own tool	ST 7426	7426	ST-6238	6238	ST-2	2	ST-121	121	ST-512	512	ST-15	15
		CGE tool		7426		6238		2		121		512		15
	#82	Own tool		7426		6238		2		121		512		15
		CGE tool		7426		6238		2		121		512		15
	#93	Own tool				6238						512		15
		CGE tool		7426		6238						512		15
	#95	Own tool												
		CGE tool		7426		6238								
	#104	Own tool		7426		6238		2		121		512		
		CGE tool		7426		6238		2		121		512		15
	#114*	Own tool												
		CGE tool		**Unk**		**Unk**								15
	#115	Own tool		7426		6238		2		121				
		CGE tool		7426		6238		2		121				

* Laboratory #114 mixed up the two strains GMI16-001-BACT and GMI16-002-BACT as well as GMI16-001-DNA and GMI16-002-DNA why the incorrect MLSTs. Some laboratories did not report MLST data (own tool) but these were provided by PT-organizer (CGE tool) marked in light gray. Deviating results indicated in bold.

Table L.6: Determined antimicrobial resistance genes in *Campylobacter* GMI16-002 for both the bacterial culture and DNA received

	Participant		GMI16-002																			
BACT	#77	Own tool	aadE	aph(3')-III	aph(2'')-like	tet(O)-like																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#82	Own tool	aadE	aph(3')-III	aph(2'')-like	tet(O)-like	gyrA T86I (Quinolone)	23S A2075G (Macrolide)														
		CGE tool	aadE	aph(3')-III		tet(O)																
	#93	Own tool	aadE	aph(3')-III	aph(2'')-like	tet(O)-like																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#95	Own tool																				
		CGE tool	aadE	aph(3')-III		tet(O)																
	#104	Own tool	aadE	aph(3')-III		tet(O)-like																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#114	Own tool																				
		CGE tool																				
	#115	Own tool																				
		CGE tool	aadE	aph(3')-III		tet(O)																
DNA	#77	Own tool	aadE	aph(3')-III	aph(2'')	tet(O)																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#82	Own tool	aadE	aph(3')-III	aph(2'')-If-like	tet(O)-like	gyrA T86I (Quinolone)	23S A2075G (Macrolide)														
		CGE tool	aadE	aph(3')-III		tet(O)																
	#93	Own tool	aadE	aph(3')-III	aph(2'')-like	tet(O)-like																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#95	Own tool																				
		CGE tool	aadE	aph(3')-III		tet(O)																
	#104	Own tool	aadE	aph(3')-III		tet(O)-like																
		CGE tool	aadE	aph(3')-III		tet(O)																
	#114*	Own tool																				
		CGE tool					**aac(6')-IIc**	**blaSHV-12**	**blaCTX-M-15**	**blaTEM-1B**	**dfrA18**	**strB**	**strA**	**ere(A)**	**tet(A)**	**tet(D)**	**sul2**	**sul1**	**QnrB49**	**floR**		
	#115	Own tool																				
		CGE tool	aadE	aph(3')-III		tet(O)																

* Laboratory #114 mixed up the two strains GMI16-001-BACT and GMI16-002-BACT why the incorrect AMR profile. The expected AMR profile for GMI16-002-BACT was reported for the pan-susceptible GMI16-001-BACT. Similarly, the laboratory #114 mixed up the GMI16-001-DNA and GMI16-002-DNA why the incorrect AMR profile for GMI16-002-DNA. Some laboratories did not report AMR data (own tool) but these were provided by PT-organizer (CGE tool) marked in light gray. Deviating results indicated in bold.

Table L.7: Determined antimicrobial resistance genes in *Klebsiella* GMI16-005 for both the bacterial culture and DNA received

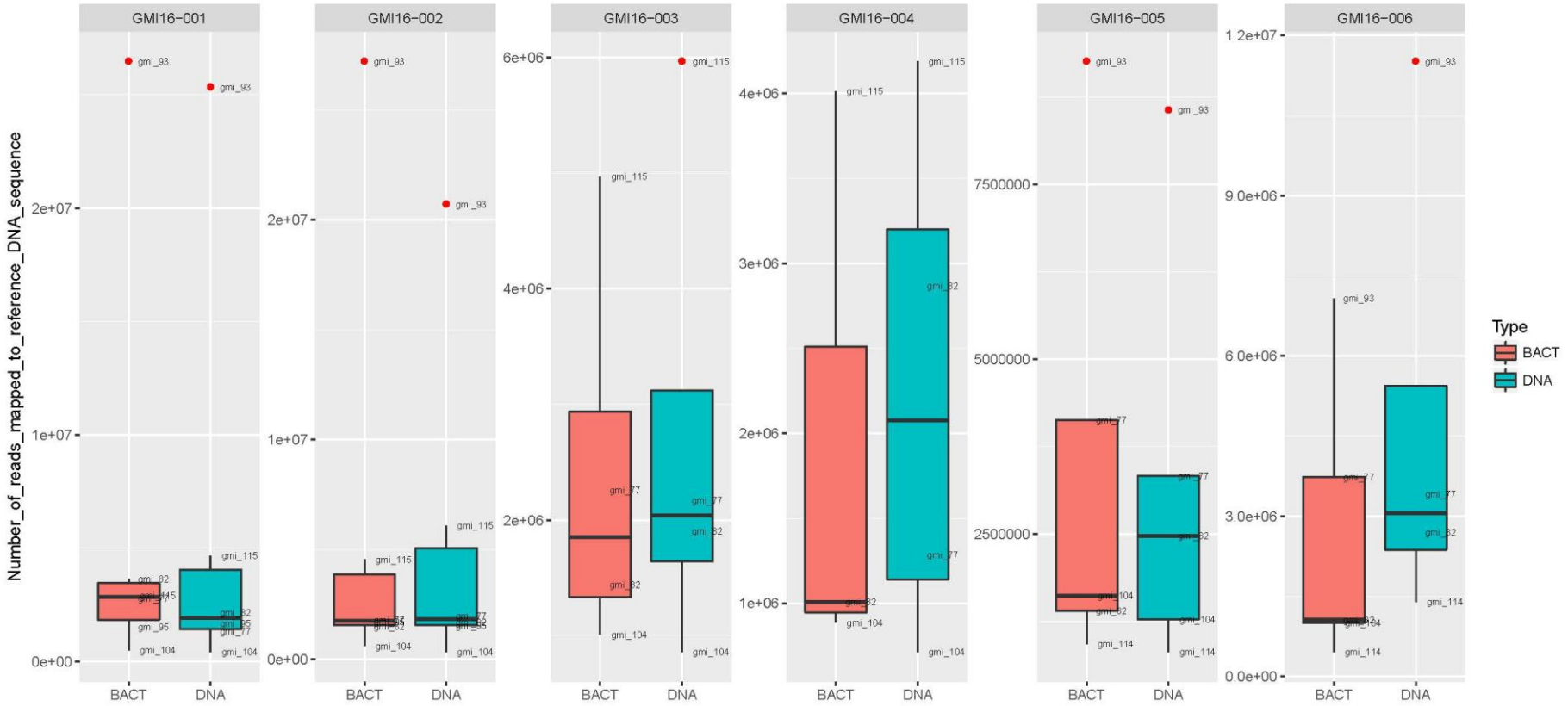
	Participant		GMI16-005													
BACT	#77	Own tool	aadA2	aac(6')-Ib)	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr	fosA	mph(A)	catA1	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#82	Own tool	aadA2-like	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB		fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#93	Own tool	aadA2	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr-like	fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#95	Own tool														
		CGE tool														
	#104	Own tool	aadA2-like	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr-like	fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#114	Own tool														
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#115	Own tool														
		CGE tool														
DNA	#77	Own tool	aadA2	aac(6')-Ib)	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr	fosA	mph(A)	catA1	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#82	Own tool	aadA2-like	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB		fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#93	Own tool	aadA2	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr-like	fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2	aac(6')-Ib	blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr		mph(A)	catA1	sul1	dfrA12
	#95	Own tool														
		CGE tool														
	#104	Own tool	aadA2-like	aac(6')-Ib	blaTEM-1A-like	blaKPC-3	blaOXA-9-like	blaSHV-11	oqxA	oqxB	aac(6')Ib-cr-like	fosA-like	mph(A)	catA1-like	sul1	dfrA12
		CGE tool	aadA2		blaTEM-1A	blaKPC-3	blaOXA-9	blaSHV-11	oqxA	oqxB			mph(A)	catA1	sul1	dfrA12
	#114	Own tool														
		CGE tool														
	#115	Own tool														
		CGE tool														

Data for the CGE tool were provided by PT-organizer and marked in light gray. Deviating results indicated in bold.

Table L.8: Determined antimicrobial resistance genes in *Klebsiella* GMI16-006 for both the bacterial culture and DNA received

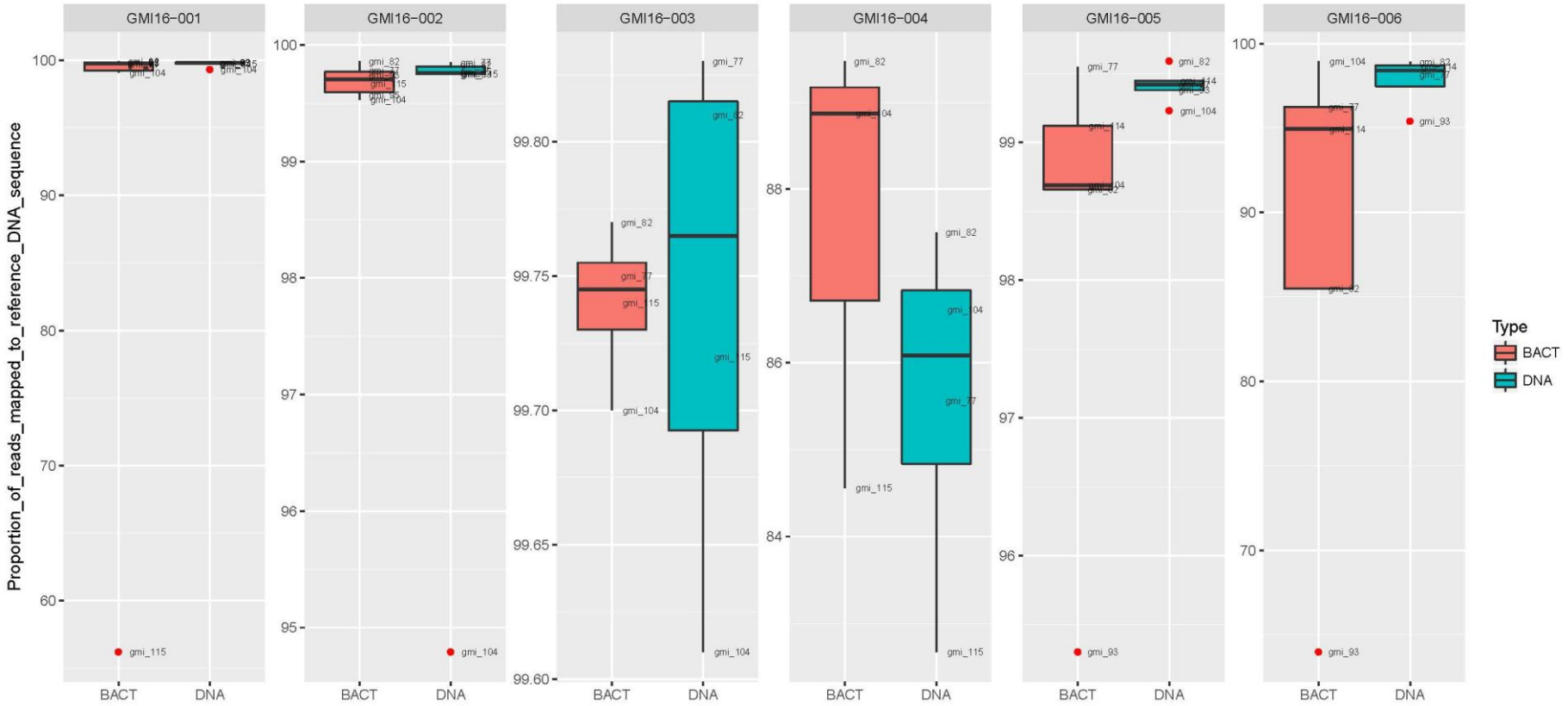
	Participant		GMI16-006																					
BACT	#77	Own tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTem-1b		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14	fosA		blaSHV-1	
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14			blaSHV-1	
	#82	Own tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')-VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B-like		QnrS1	oqxA-like	oqxB-like		sul2	tet(D)	dfrA14-like	fosA-like	blaSHV-28		
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15		blaTEM-1A	QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14		blaSHV-28		
	#93	Own tool	aadA1	aac(6')-Ib		aph(3')VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9-like	blaCTX-M-15	blaTEM-1B		QnrS1	oqxA-like	oqxB-like	aac(6')Ib-cr-like	sul2	tet(D)	dfrA14-like	fosA-like		blaSHV-1	
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14				
	#95	Own tool																						
		CGE tool																						
	#104	Own tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')-VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9-like	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB-like	oqxA-like	aac(6')Ib-cr-like	sul2	tet(D)	dfrA14-like,	fosA	blaSHV-28		
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15		blaTEM-1A	QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14				
	#114	Own tool																						
		CGE tool																						
	#115	Own tool																						
		CGE tool																						
DNA	#77	Own tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTem-1b		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14	fosA		blaSHV-1	
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14				
	#82	Own tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')-VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B-like		QnrS1	oqxA-like	oqxB-like		sul2	tet(D)	dfrA14-like	fosA-like	blaSHV-28		
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14		blaSHV-28		
	#93	Own tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9-like	blaCTX-M-15	blaTEM-1B		QnrS1	oqxA-like	oqxB-like	aac(6')Ib-cr-like	sul2	tet(D)	dfrA14-like	fosA-like		blaSHV-1	blaLEN12-like
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14				
	#95	Own tool																						
		CGE tool																						
	#104	Own tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')-VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9-like	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB-like	oqxA-like	aac(6')Ib-cr-like	sul2	tet(D)	dfrA14-like,	fosA	blaSHV-28		
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId-like	aph(3')-VIa-like	strA-like	strB-like	blaNDM-1	blaOXA-9-like	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB-like	oqxA-like	aac(6')Ib-cr-like	sul2	tet(D)	dfrA14-like,	fosA	blaSHV-28		
	#114	Own tool																						
		CGE tool	aadA1	aac(6')-Ib	aac(3)-IId	aph(3')-VIa	strA	strB	blaNDM-1	blaOXA-9	blaCTX-M-15	blaTEM-1B		QnrS1	oqxB	oqxA	aac(6')Ib-cr	sul2	tet(D)	dfrA14				
	#115	Own tool																						
		CGE tool																						

Data for the CGE tool were provided by PT-organizer and marked in light gray. Deviating results indicated in bold.



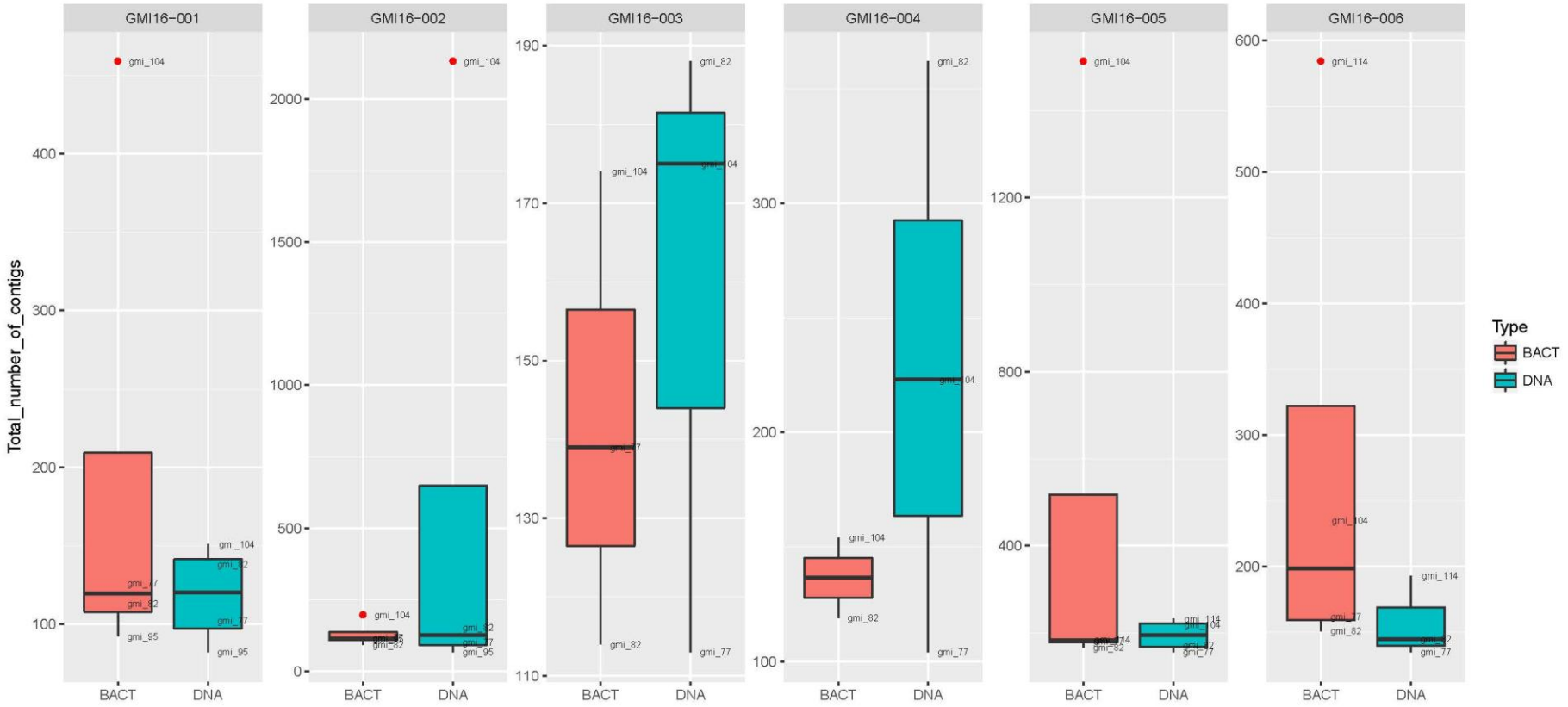
The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.
Results for participant 2 omitted for both sample types of strain 1, 4, 5, and 6. The whiskers represent minimum and maximum values (range) and the box represent the Q1, Median, and Q3, respectively.

Figure L.1: Number (No) of reads mapped to the reference sequence



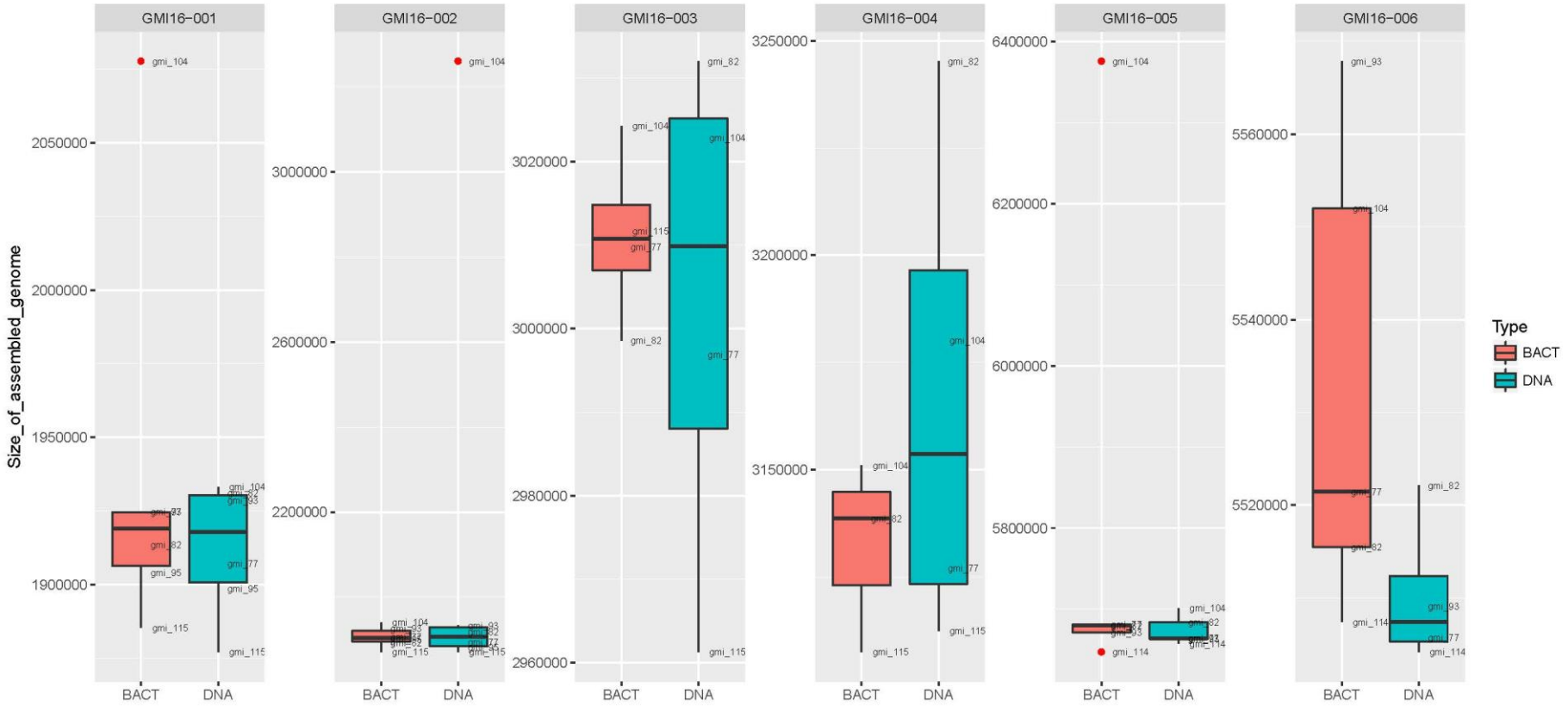
The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.2: Proportion (%) of reads mapped to the reference DNA sequence



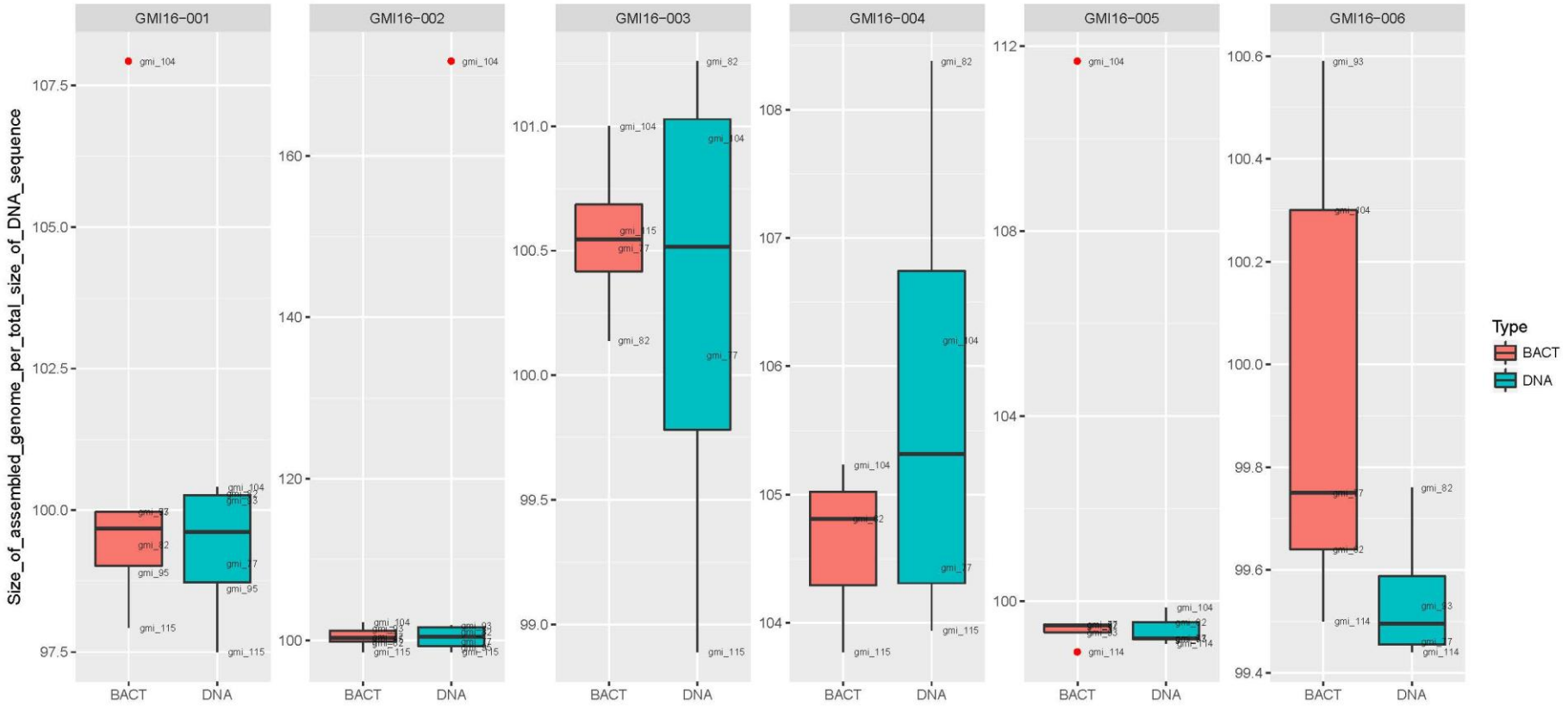
The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.3: Total number (No) of contigs



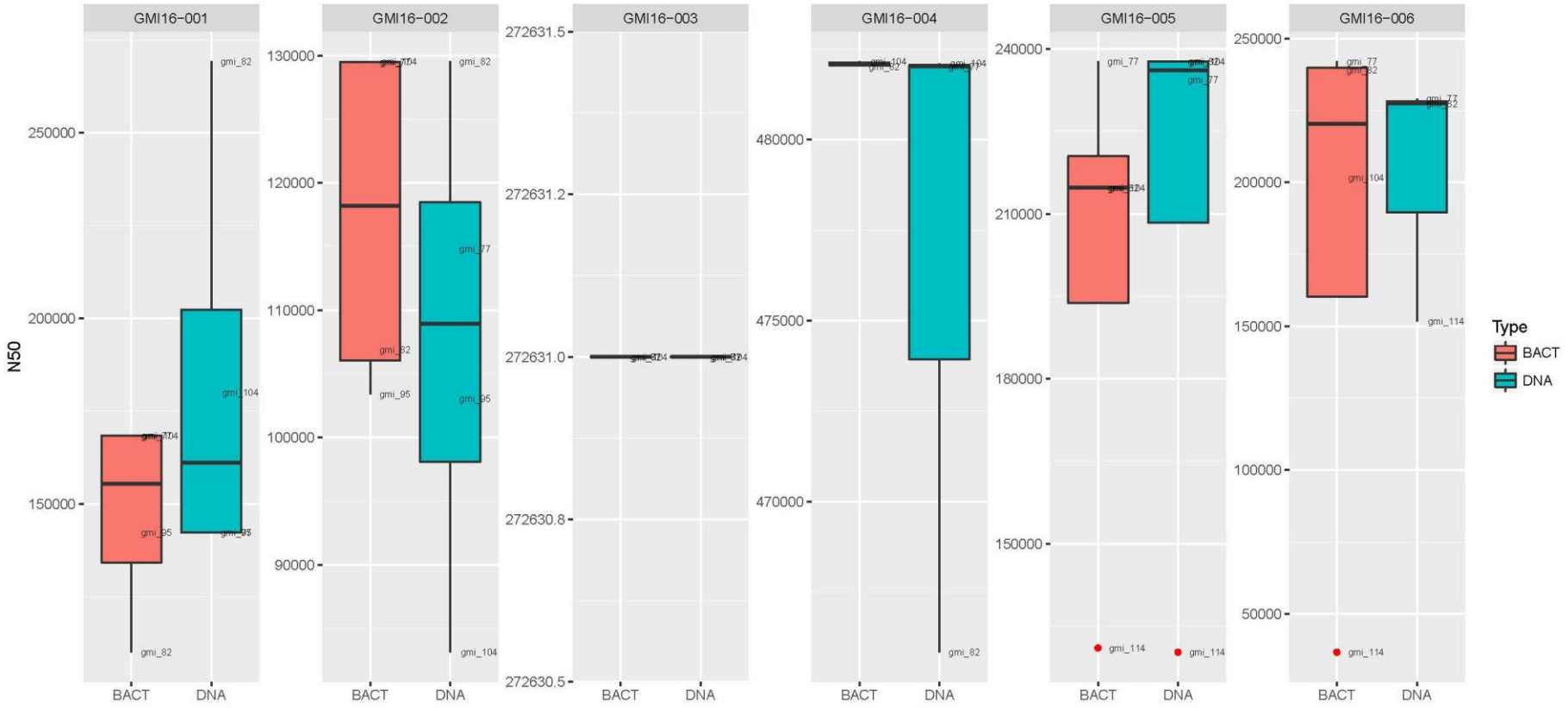
The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.4: Size (bp) of the assembled genome



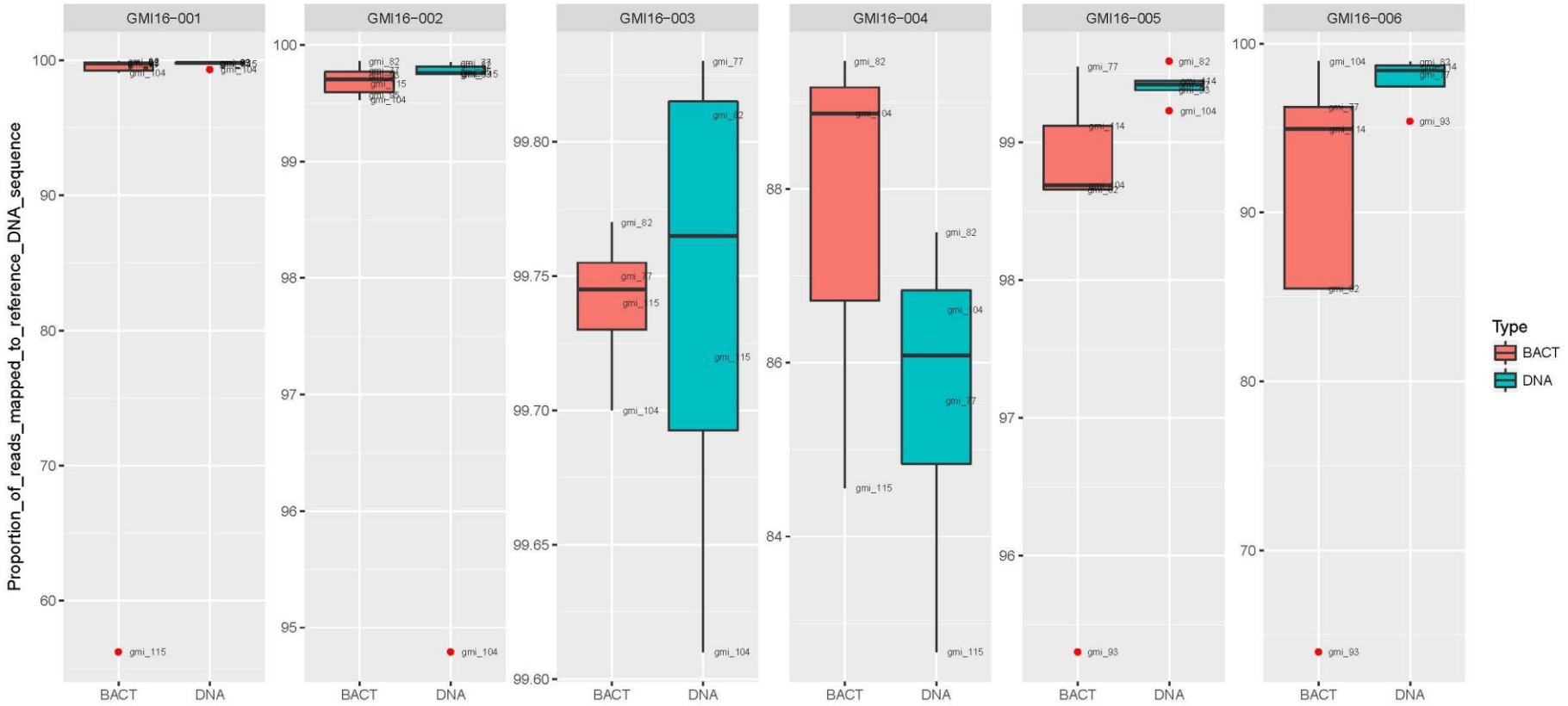
The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.5: Proportion (%) of the assembled genome per reference DNA sequence



The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.6: N50 - average length (bp) of sequences



The black line inserted the box represent the median which indicates 50% of the data being greater than this value. The top and botton of the box indicate the upper and lower quartiles which is 25% of the data being greater or lower than this value. The end of the whiskers indicates the maximum (greatest value) or minimum (lowest value) excluding outliers. The red dots represent values considered outliers.

Figure L.7: Depth of coverage (X) of the sequences

Appendix M – The ENGAGE Proficiency Test Report 2017

THE ENGAGE PROFICIENCY TEST REPORT 2017

Oksana Lukjancenko, Susanne Karlsmose Pedersen, Pimlapas Leekitcharoenphon, Martin Christen Frølund Thomsen, Lukasz Dariusz Dynowski, Jose Luis Bellod Cisneros, Ole Lund, Rene S. Hendriksen

2nd edition, January 2018

Technical University of Denmark, National Food Institute, Research Group of Genomic Epidemiology, Kgs. Lyngby, Denmark

Contents

List of Abbreviations

AMR	Antimicrobial resistance
BWA	Burrows-Wheeler Aligner
CC	Clonal Complex
CGE	Center for Genomic Epidemiology
CIA	Critically Important Antimicrobials
DNA	Deoxyribonucleic acid
DTU	Technical University of Denmark
E. coli	*Escherichia coli*
GMI	Global Microbial Identifier
HGAP	Hierarchical Genome Assembly Process
IATA	International Air Transportation Association
MLST	Multi Locus Sequence Typing
PT	Proficiency test
QC	Quality Control
S. aureus	*Staphylococcus aureus*
SMRT	Single-molecule real-time
SNP	Single Nucleotide Polymorphism
US FDA	U.S. Food and Drug Administration
WGS	Whole Genome Sequencing

1. Introduction

The main objective of this proficiency test (PT) is to facilitate the production of reliable laboratory results of consistently good quality within the area of whole genome sequencing (WGS).

The PT evaluates the consistency and robustness of ENGAGE consortium members' ability to perform deoxyribonucleic acid (DNA) extraction, library preparation, the WGS, and assembly following different laboratory protocols, software tools, and sequence platforms for the reliability of submitted sequence data to the public repositories. This ensures harmonization and standardization in WGS and data analysis, with the aim to produce comparable data for the ENGAGE initiative. To meet these objectives, the laboratory work and analyses for this PT were performed using the methods routinely employed in the individual laboratories.

The PT consists of a "wet-lab" component targeting three common bacterial pathogens. The wet-lab component assesses the laboratories' ability to perform DNA preparation, sequencing procedures and, if laboratories routinely do so, the analysis of epidemiological markers, i.e. Multi Locus Sequence Typing (MLST) and antimicrobial resistance (AMR) genes.

The individual laboratory data are confidential and only known by the participating laboratory and the PT organizers (DTU Food).

Materials and Methods

Participating laboratories

A pre-notification to announce the ENGAGE proficiency test was distributed on the 8 May 2017 by e-mail to the eight ENGAGE consortium partners. Seven of the eight partners signed up and participated in the PT. The National Institute of Public Health - National Institute of Hygiene, Poland, did not participate as they have not initiated in-house WGS.

Strains

In 2017, two strains each of *Salmonella* (GMI17-001, GMI17-002) *Escherichia coli* (*E. coli*) (GMI17-003, GMI17-004), and *Staphylococcus aureus* (*S. aureus*) (GMI17-005, GMI17-006) were selected for the wet-lab component. In a GMI end-user analysis of what species to target, *Salmonella* and *E. coli* were indicated to be of most interest to the end-users (Moran-Gilad et al., 2015). *S. aureus*, a public health relevant pathogen carrying a significant burden worldwide, was also indicated as important and therefore selected for this PT. All of the selected strains are resistant to critical important antimicrobials (CIAs) for human health. *Salmonella*, GMI17-001 belong to *S.* Bovismorbificans/*S.* Hindmarsh and confer resistance to colistin and harbor the *mcr*-1 gene whereas GMI17-002 is *S.* Westhampton and a carbapenemase producer containing both $bla_{\text{NDM-1}}$ and $bla_{\text{CTX-M-15}}$. The two *E. coli* strains, GMI17-003, GMI17-004 are also resistant to CIAs harboring the $bla_{\text{NDM-7}}$ and *mcr*-1 genes, respectively. Finally, *S. aureus* GMI17-005 was included, and being a livestock associated (LA)-MRSA, *spa* type t034 of a novel sequence type (ST) which is a double-locus variant of ST130 belonging to Clonal Complex (CC) CC398 and additionally resistant to linezolid due to the presence of the *cfr*-gene. The second *S. aureus* GMI17-006 is a LA-MRSA *spa* type t843, ST130, belonging to CC130.

Individual sets of the test strains were produced as agar stab cultures (nutrient) and the corresponding DNA was purified and pooled by DTU Food prior to distribution in individual vials for each participating laboratory.

To better enable the assessment of the differences in the sequences generated by the laboratories, each of the six strains in the wet-lab component was sequenced on the PacBio by the Food and Drug Administration in the United States of America (US FDA) to obtain a closed reference genome. Initially, 10 kb template libraries were created using "10 kb DNA Template Prep Kit 1.0" from Pacific Biosciences. Subsequently, the libraries were sequenced using C2 chemistry on single-molecule real-time (SMRT) cells with a 180 min collection protocol. The data were *de novo* assembled using the Hierarchical Genome Assembly Process

(HGAP) within the Pacific Biosciences SMRTAnalysis software package. Polishing and finishing the genome were performed with custom python scripts, Quiver and Gepard, a dot plot tool to identify overlapping regions. The following reference genomes were generated for the PT, *Salmonella* (GMI17-001), *Salmonella* (GMI17-002), *E. coli* (GMI17-003), *E. coli* (GMI17-004), *S. aureus* (GMI17-005), and *S. aureus* (GMI17-006).

Distribution

On 21 August 2017, bacterial strains in agar stab cultures together with the corresponding purified and dried DNA and a welcome letter were dispatched in double pack containers (class UN 6.2) to the participating laboratories according to the International Air Transport Association (IATA) regulations as UN3373, biological substances Category B.

Procedure

The protocol was made available on the website allowing the PT laboratories access to all necessary information at any time (http://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-test-2017). Additional relevant information was distributed by email directly to the laboratories.

The protocol presented instructions as to the handling of the received bacterial cultures and DNA.

Laboratories were requested to capture information in relation to the questions presented in the online survey.

This report summarizes the results and allows for ensures full anonymity for the laboratories, as only the PT-organizers has access to the individual results.

Sequencing protocols and quality metrics

2.4.1. Online survey of the sequencing capabilities

Apart from three questions relating to the contact information of the laboratory, 40 questions focused on the storage of bacterial cultures and DNA prior to analysis, the cultivation and DNA extraction procedure, the quality assurance parameters applied and also on the details related to the sequencing and analysis of the obtained sequencing data, were asked.

The laboratories submitted raw sequence files in fastq format. As part of the analysis, the reads were *de novo* assembled using the SPAdes v 3.6.1 software. Reads were aligned to reference chromosomes and plasmids using Burrows-Wheeler Aligner (BWA)-MEM algorithm with default settings. Samtools was used to filter the reads that did not map. MLST genes and ST were predicted using MLST tool provided by Center for Genomic Epidemiology (CGE) (https://cge.cbs.dtu.dk/services/MLST/) Larsen et al., 2012). Antimicrobial resistance genes were predicted using ResFinder tool https://cge.cbs.dtu.dk//services/ResFinder/database.php (Zankari et al., 2012).

For the raw reads, the following QC metrics were calculated:

- Numbers of reads (for paired-end reads, the total numbers of reads is calculated as the sum of reads in the two files)
- Numbers of reads after trimming
- Numbers of unmapped reads
- Number of reads that map to the total reference DNA (chromosome + any plasmids) using BWA
- Number of reads that map to reference chromosome
- Number of reads that map to the reference plasmid #1 - #4
- Proportion (%) of reads that map to reference chromosome out of all sequence reads in the sample.

- Coverage of the reference chromosome (fraction of chromosome positions that were covered by at least one read pair).
- Coverage of the reference plasmid #1 - #4 (number of plasmid positions that were covered by at least one read pair).
- Depth of coverage of total DNA
- Depth of coverage of the reference chromosome
- Depth of coverage of the reference plasmid #1 - #4
- Q-score R1 (Base calling accuracy, Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencer).
- Q-score R2

For the assemblies, the following QC parameters were calculated:

- Total size of assembly (bp) (all contigs)
- Proportion (%) of size of assembly that map to the total size of DNA
- Total number of contigs
- Number of contigs with a length above 200 bp
- N50 (defined as the length of the shortest contig, in the set of largest contigs that represents at least 50% of the assembly)

In addition to the calculation of the above QC metrics and parameters, laboratories were requested to provide the identification of the corresponding MLST and AMR genes of the strains to support the assessment of the sequence quality. Laboratories identified the MLSTs and AMR genes using the software of their choice. To assess the proficiency of the laboratories, the PT organizers used a command line version of the CGE MLST-Finder v.1.8 and ResFinder 3.0 (Threshold for %ID = 98% and HSP/Query length = 60%) including the CGE standard assembly pipeline on the laboratories raw reads to compare the results with those reported by the laboratories. Furthermore, strain-specific reference routed phylogenetic single nucleotide polymorphism (SNP) trees were created using the raw reads of both the culture and corresponding DNA submitted by each of the laboratories. This will support the assessment of the sequence quality of the laboratories.

Phylogenetic SNP trees were created using the pipeline; CSI phylogeny v.1.4 available from CGE. The paired-end reads were mapped to the reference genomes; using BWA version 0.7.2. The depth at each mapped position was calculated using genomeCoverageBed, which is part of BEDTools version 2.16.2. SNPs were called using 'mpileup' module in SAMTools version 0.1.18. SNPs were filtered out if the depth at the SNP position was not at least 10X or at least 10% of the average depth for the particular genome mapping. Subsequently, SNPs were selected when meeting the following criteria: 1) a minimum distance of 10 bp between each SNP, 2) the mapping quality was more above 25, 3) the SNP quality was more than 30 and 4) all indels were excluded.

The qualified SNPs from each genome were concatenated to a single alignment corresponding to position of the reference genome. The concatenated sequences were subjected to maximum likelihood tree using FastTree (Price et al., 2010).

Results

Participation

Seven laboratories responded to the pre-notification and were enrolled in the ENGAGE PT. When the deadline for submitting results was reached, all seven laboratories had uploaded data. All seven partners, #141, #146, #152, #156, #170, #182, and #185 submitted raw reads obtained from both the culture and the DNA for both strains of the three bacterial species with the exception of laboratory #152, which only participated in the *E. coli* and *S. aureus* trials.

Method description

Three laboratories immediately initiated processing the cultures after receipt. The other four laboratories stored the cultures at 4ºC. Similarly, the DNA was stored at 4ºC by three laboratories whereas four laboratories stored the DNA samples at room temperature. Only one laboratory processed the DNA upon receipt of the samples. All the laboratories incubated the bacterial strains at 37˚C ranging from 16 h to 24 h.

The laboratories also reported the DNA extraction procedures which indicated a high degree of variation among the kits being used.

Among the seven laboratories, two laboratories used an automatic extraction based on the following instruments and kits, laboratory #182, QIAsymphony DSP DNA minikit added RNAse at 37ºC for 15 minutes and elution in water rather than buffer for Gram-negative bacteria and pre-lysis steps to include incubation in presence of lysozyme, lysostaphin, proteinase K for Gram positive bacteria and laboratory #185, King-Fisher DUO Prime, MagMAX Core Nucleic Acid Purification kit with a sample volume of 270 µL. The other five laboratories followed a manual extraction protocol for the three bacterial species, Easy - DNA (#141), PureLink™ Genomic DNA Mini Kit included proteinase K digestion and lysostaphin for *S. aureus* (#146), QIAamp DNA Mini Kit (#152/ #156) and Genomic mini included lysostaphin digestion for the Gram-positive bacteria (#170).

In addition, the laboratories also reported the DNA concentration (ng/µL) and DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation. The DNA concentration (ng/µL) prior to library preparation was measured on Qubit by the five laboratories using a manual extraction procedure whereas laboratory #182 used the Quantifluor dsDNA system and laboratory #185 used GloMax® 96 Microplate Luminometer, an automated plate reader using fluorescence. The DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation was measured by three laboratories on Nanodrop (#146, #152, and #170) whereas the remaining four laboratories did not measure the quality. In addition, two laboratories, #146 and #170 performed quality check to verify the quality of the DNA on a gel.

All of the laboratories used commercial kits for library preparation and all related to the used sequencing platform, Nextera XT DNA Library Preparation Kits. Modifications to the protocols were reported by laboratory #146 that used half the volume of the reagents and laboratory #152 which for manual normalization of purified libraries used TrisHCl 10 mM.

The genomic DNA was prepared for pair-end sequencing by all seven laboratories. The libraries were sequenced using the following read length and platform, Illumina MiSeq by four laboratories (#146, #152, #156, (251 bp) and #170 (300 bp)), Illumina NextSeq by laboratory #141 (read length not known), Illumina HiSeq 2500 by laboratory #185 (200 bp), and Genome Sequencer Junior System (454) by laboratory #182 (150 bp).

Five laboratories indicated that if assembled by themselves, they would have used FoodQCPipeline (trimmed by bbduk2 (part of the suit bbtools version 36.49) and *de novo* assembly by SPAdes) (#141), Velvet Assembler 1.3 (#156), and SPAdes 3.9.0 (#146, #170, #185).

Sequencing – quality markers

All the seven laboratories participating in the PT trial submitted sequencing data for all the six test strains, except laboratory #152 which did not submit sequencing data for the *Salmonella* GMI17-001 and GMI17-002 from both the bacterial culture and corresponding DNA.

The sequencing quality of all submitted data was evaluated for potential contamination or a low performance by accessing the above quality parameters.

3.2.1. *Salmonella* – genome size

For the *Salmonella* genomes, GMI17-001-BACT and GMI17-001-DNA with an expected genome size of 4,687,697 bp, none of the laboratories were considered outliers as for the obtained size (Figures M.1-M.2).

For the second *Salmonella* genome, GMI17-002-BACT with a genome size of 5,119,002 bp, none of the laboratories were considered outliers (Figures M.1-M.2). Laboratory #71, however, provided the highest assembled genome size of the sample, GMI17-002-DNA with 5,219,504 bp compared with a genome size of 5,119,002 bp (102 %) and was considered an outlier.

3.2.2. *Salmonella* – Q-score

All laboratories obtained a base calling accuracy, Phred quality score (Q score), between Q30 and Q40 for all *Salmonella* genomes and sample types indicating a 99.9 % base call accuracy with the probability of an incorrect base call in 1 out of 1,000 bp (Figures M.3-M.4).

3.2.3. *Salmonella* – reads and coverage

Regarding the number of reads produced, laboratory #141 was considered an outlier for both *Salmonella* genomes, GMI17-001-BACT, GMI17-002-BACT and GMI17-002-DNA, reaching 10,589,514 bp, 10,539,188 bp and 9,105,498 bp, respectively (Figure M.5). None of the laboratories reported high amounts of unmapped reads for the *Salmonella* genomes, GMI17-001 and GMI17-002-BACT and thus no outliers were identified (Figure M.10). Some consistency was observed in relation to underperformance when analyzing the reported number of reads mapping to the reference DNA, the reference chromosome, after trimming, and plasmid 1 (Figures M.6-M.9, M.11). Laboratory #141 reported the highest number of reads mapping to the reference DNA sequence for both sample types and for both *Salmonella* genomes e.g. 9,180,225 bp for GMI17-001-BACT and was considered as an outlier (Figure M.7). Interestingly, laboratory #185 obtained a very low proportion (64.9%) of reads mapping to the reference DNA sequence for GMI17-002-DNA and was identified as an outlier (Figure M.8).

All laboratories obtained an almost 100% in the proportion of reads mapping to the reference DNA sequence and to plasmid 1 (Figures M.8, M.11).

Laboratory #141 that was considered as an outlier for mapping reads to the reference DNA also sequenced the genomes including the chromosome and plasmid with a high sequencing depth of e.g. 229X for GMI17-002-DNA and was considered an outlier compared to the other laboratories in relation to sequencing depth (Figure 15-17).

The coverage of the reference chromosome and plasmid 1 for both sample types of the two *Salmonella* genomes were quite high with almost all laboratories close to having a coverage of 100% (Figures M.21-M.22).

The average insert sizes of the *Salmonella* genomes of both sample types seemed in many cases to be a bit greater than the read length except for laboratories #146 and #182 for which the insert sizes were approximately the double of the read length with #146 considered an outlier (Figure M.26).

3.2.4. *Salmonella* – contigs and N50

The total number of contigs and number of contigs above 200 bp were estimated revealing no outliers (Figure M.27). Similarly to the contigs, also the N50 was estimated based on the submitted sequence data and revealed no outliers. The lowest observed N50 for *Salmonella* GMI17-001 was 177,550 bp and was obtained by laboratory #156 for the DNA sample (Figures M.29-M.30). Similarly, the same laboratory, #156 obtained the lowest N50 value of 82,918 bp for the DNA sample of *Salmonella* GMI17-002 (Figure M.29).

3.2.5. *E. coli* – genome size

For the *E. coli* genome, GMI17-003 with a genome size of 4,923,235 bp, laboratory #156 was the only laboratory considered an outlier with a low assembled genome size of 2,397,310 bp (43.6 %) for the BACT sample (Figures M.1-M.2). No outliers were observed for the corresponding DNA sample. Regarding the proportion of reads mapped, laboratory #156 was considered an outlier with a slightly lower assembled genome size of 4,946,278 bp (99.6 %) compared to the expected genome size of 4,923,235 bp. In addition, laboratory #141 was also considered an outlier but with a slightly higher assembled genome size of 4,989,924 bp (100.5 %) compared to the reference (Figures M.1-M.2).

3.2.6. *E. coli* – Q-score

All laboratories obtained a base calling accuracy, Phred quality score (Q score) ranging between Q28.9 (laboratory #185, GMI17-003-BACT) and Q37.2 (laboratory #182, GMI17-004-DNA) for all *E. coli* genomes and sample types indicating an almost 99.9 % base call accuracy with the probability of an incorrect base call in 1 out of 1,000 bp for most laboratories (Figures M.3-M.4).

3.2.7. *E. coli* – reads and coverage

Laboratory #141 was considered outlier measuring the number of reads produced for both samples types of the *E. coli* genomes, GMI17-003 and GMI17-004 reaching 11,781,918 bp (GMI17-003-BACT), 11,261,596 bp (GMI17-003-DNA), 9,194,922 bp (GMI17-004-BACT), and 12,787,574 bp (GMI17-004-DNA), respectively (Figure M.5). The same laboratory, #141, revealed a high sequencing depth for both *E. coli* genomes including plasmid 1 to 4 and both sample types but considered an outlier for the DNA sample type e.g. 353X for GMI17-004-DNA when compared to the other laboratories (Figures M.15-M.20). The laboratory which reported the highest amount of unmapped reads for the *E. coli* genomes, GMI17-003-DNA and GMI17-004-DNA was laboratory #185 with 1,142,545 bp and 1,079,169 bp, respectively, and was therefore considered an outlier (Figure M.10). Some consistency in the laboratory's underperforming was observed for the reporting the number of reads mapping to the reference DNA sequence, the reference chromosome, and after trimming (Figures M.6-M.9, M.11-M.14). Laboratory #156 reported the lowest number and proportion of reads mapping to the reference chromosome for the GMI17-003-BACT with 44,333 bp (4.1%) and was therefore considered as an outlier (Figure M.9). Similarly, laboratory #185 reported low numbers and proportions of reads mapping to the reference chromosome of both *E. coli* genomes, GMI17-003 and GMI17-004 and both samples types e.g. 1,421,916 bp (53.2%) for GMI17-003-BACT (Figure M.9). Laboratory #141 reported the highest number and proportion of reads mapping to the reference chromosome and plasmid 1 to 4 for both samples types of the *E. coli* of both genomes e.g. 11,832,263 bp (95.8%) for GMI17-004-DNA, and was therefore considered as an outlier (Figures M.9, M.11-M.14). In general, the proportion of reads mapping to the reference DNA sequence were lower than 100% for all of the laboratories (Figure M.8). The coverage to the reference chromosome and plasmid 1 to 4 for both sample types of both *E. coli* genomes were quite high with almost all laboratories presenting coverages of 100%, except for laboratory #156 which obtained e.g. a coverage to the reference chromosome of 67.1% for GMI17-003-BACT (Figures M.21-M.25).

The average insert sizes of the *E. coli* genomes of both sample types seemed to be similar to *Salmonella* with laboratory #146 exhibiting an insert size approximately the double of the read length and laboratory #146 was therefore considered an outlier (Figure M.26).

3.2.8. *E. coli* – contigs and N50

The total number of contigs and the number of contigs above 200 bp were estimated and a number of laboratories were considered outliers including laboratory #156 and #141 for the *E. coli* genomes, GMI17-003-BACT and GMI17-004-BACT, respectively (Figures M.27-M.28). Laboratory #156 obtained an exceptionally high number of contigs above 200 bp of 3,397. Similarly to the contigs, also the N50 was estimated based on the submitted sequence data. Laboratory #156 was also considered an outlier related to the N50 for GMI17-003-BACT with 735 bp (Figure M.29).

3.2.9. *S. aureus* – genome size

For the *S. aureus* genomes, GMI17-005 and GMI17-006 with the genome sizes of 2,790,673 bp, and 2,928,160 bp, respectively, laboratory #156 was the only laboratory considered an outlier with a too large size of the assembled genome of 3,017,011 bp (108.1%) and 3,283,558 bp (112.1%) for the BACT samples (Figures M.1-M.2). No outliers were observed for the corresponding DNA sample.

3.2.10. *S. aureus* – Q-score

All laboratories obtained a base calling accuracy, Phred quality score (Q score) ranging between Q31.3 (laboratory #170, GMI17-005-BACT) and Q38.6 (laboratory #186, GMI17-005-DNA) for all *S. aureus* genomes and sample types indicating an almost 99.9 % base call accuracy with the probability of an incorrect base call in 1 out of 1,000 bp for most laboratories (Figures M.3-M.4).

3.2.11. *S. aureus* – reads and coverage

Laboratory #141 were considered an outlier related to the number of reads produced for both sample types of the *S. aureus* genomes, GMI17-005 and GMI17-006 reaching 11,198,996 bp (GMI17-005-DNA), and 9,792,850 bp (GMI17-006-DNA), respectively (Figure M.5). The same laboratory, #141, revealed a high sequencing depth, 421.94X for the *S. aureus* genome, GMI17-006-DNA including plasmid 1 (2088.57X) and was considered an outlier when compared to the other laboratories (Figures M.15-M.17). The laboratory which reported the highest amount of unmapped reads for the *S. aureus* genome, GMI17-006-DNA, was also laboratory #141 with 938,292 bp and this laboratory was considered an outlier (Figure M.10). Some consistency to the above parameters was observed in relation to the reported number of reads mapping to the reference DNA sequence, the reference chromosome, and after trimming (Figures M.6-M.9, M.11). Laboratory #141 reported the highest number and proportion of reads mapping to the reference chromosome for both samples types of the *S. aureus* of both genomes e.g. 9,403,511 bp (94.0%) for GMI17-005-BACT and for plasmid 1 for genome GMI17-006 and was therefore considered as an outlier (Figure M.9). In general, the proportion of reads mapping to the reference DNA sequence was lower than 100% for all of the laboratories (Figure M.8). The coverage to the reference chromosome and plasmid 1 for both samples types of the *S. aureus* of both genomes were quite high with almost all laboratories having coverage of 100% (Figures M.21-M.22).

The average insert sizes of the *S. aureus* genomes of both sample types seemed to be similar as for *Salmonella* with laboratory #146 having an insert size approximately the double of the read length and laboratory #146 was considered an outlier (Figure M.26).

3.2.12. *S. aureus* – contigs and N50

The total number of contigs and the number of contigs above 200 bp were estimated and none of the laboratories were considered outliers (Figures M.27-M.28). Similarly to the contigs, also the N50 was

estimated based on the submitted sequence data and none of the laboratories were considered being outliers (Figure M.29).

Sequencing – MLST, and antimicrobial resistance genes

For *Salmonella* GMI17-001, the expected MLST was ST142 which was identified by almost all of the laboratories except for laboratory #141 which identified the genome GMI17-001 of both sample types as ST377 (Table M.1). Laboratory #182 did not report MLST data for neither of the *Salmonella, E. coli* and *S. aureus* genomes (own tool) but these were provided by PT-organizer (CGE tool) and identified the correct MLST for GMI17-001, GMI17-003, GMI17-004 and GMI17-005 whereas it was incorrect for genome GMI17-002 (which was also the case for all other laboratories) and for GMI17-006 (as expected).

Salmonella spp.

All laboratories, except for laboratory #158 which did not submit any *Salmonella* data, detected (own tools) and reported the following expected resistance genes, $bla_{\text{TEM-1}}$, *sul*2, *tet*A, *mcr*-1, and *drf*A14 in both sample types from *Salmonella*; GMI17-001 (Table M.2). The genome also contained the genes, *aph*(6)-Id and *aph*(3")-Ib which were not reported by the laboratories but only detected using the reference CGE tool, ResFinder. Thus, one laboratory (#185) was an exception correctly reporting the gene, *aac*(6')-Iy but also incorrectly *aph*(6)-Id. In addition, all laboratories identified the genes *str*A and *str*B which according to the reference method were not present. The expected *mcr*-1 gene was not detectable in the reference GMI17-001. The contig that contained the *mcr*-1 gene was 26,432 bp large which is part of plasmid. The reason was that the reference strain was sequenced by PacBio and PacBio only sequenced part of this plasmid and missed the *mcr*-1 gene. The reason was most likely that the plasmid was lost in culturing or DNA purification.

The MLST ST14 was expected for the *Salmonella* genome; GMI17-002 which all laboratories reported correctly (Table M.1).

All of the laboratories were able, using their own and the CGE tools to identify in the *Salmonella* genome, GMI17-002, the following expected genes: $bla_{\text{CTX-M-15}}$, *aac*(6')-Ib-cr, $bla_{\text{OXA-1}}$, $bla_{\text{OXA-9}}$, $bla_{\text{OXA-10}}$, *cml*A1, *mph*(A), *cat*B4, *arr3*, and *sul1* with the exception of laboratory #182 which did not report any genes and laboratory #158 detecting *cat*B3 (Table M.3). In addition, all laboratories identified $bla_{\text{DHA-1}}$ which was not one of the expected genes present in the reference genome. Moreover, all laboratories identified $bla_{\text{NDM-1}}$ which appears not to have been captured by the PacBio sequencing of the reference. The expected $bla_{\text{NDM-1}}$ gene was not detectable in the reference GMI17-002. The $bla_{\text{NDM-1}}$ gene is located on a 4277 bp plasmid. The assembler form PacBio missed it. The reason was that the reference strain was sequenced by PacBio and PacBio has a limitation to sequence small plasmid. Thus the plasmid containing $bla_{\text{NDM-1}}$ was not sequenced completely by PacBio and further analysis is ongoing. The *Salmonella* genome GMI17-002 was also expected to contain the gene $bla_{\text{TEM-1A}}$ which most laboratories were able to detect or the variant $bla_{\text{TEM-1B}}$ except for laboratory #185 using own tools identifying the gene $bla_{\text{TEM-191}}$. Similarly, the reference was expected to contain the gene *aad*A1 which were detected by both own and CGE tools by laboratory #146. The data submitted by laboratories, #141 and #156 reported the presence of the variant *aad*A24 (own tools used). The identification of the expected gene, *aac*(3)-IIa caused issues but all laboratories were able to identify various variants of this gene.

E. coli

Almost all laboratories reported the correct MLST, ST448, for the *E. coli* genome GMI17-003 except for laboratory #156 using the reference tool, CGE tool, revealing an incorrect unknown ST (Table M.1).

Four of the laboratories provided data using own and CGE tools whereas the laboratories, #152 (both samples types), #156 (DNA sample), and #182 for which only data from the CGE tools were generated (Table M.4). All laboratories were able to identify in *E. coli* genome GMI17-003 the expected antimicrobial resistance genes, *aad*A5, *aac*(3)-IId, *aac*(6')-Ib-cr, $bla_{\text{TEM-1B}}$, *cat*A1, $bla_{\text{NDM-7}}$, $bla_{\text{OXA-1}}$, *mph*(A), *cat*B3, and *dfr*A17 except for the GMI17-003-BACT sample of laboratory #156. Laboratory #156 only detected the $bla_{\text{TEM-1B}}$ and *cat*A1 in the submitted genome of GMI17-003-BACT using the CGE tool. Furthermore, the *sul*1

gene was also expected in GMI17-003 and observed using the CGE tool by all laboratories except as mentioned in the BACT sample from laboratory #156. In addition, laboratories #170 and #185 did not report the *sul*1 gene using own tools whereas laboratory #185 also reported the presence of *mdf*(A) based on their own tool which contained this gene in the database of AMR genes excepted.

In all cases, the laboratories managed to identify the correct and expected MLST ST10 (adk-10, fumc-11, gyrb-4, icd-8, mdh-8, pura-8, reca-2) of *E. coli* genome, GMI17-004 using the reference CGE tool (Table M.1). However, using their own tool, the MLST was misclassified by four laboratories, #141, #146, #170, and #185 which reported an allele difference compared to ST10 in icd-8 (#141), whereas the rest of the four laboratories reported a novel ST.

Five of the laboratories provided data using own and CGE tools whereas the laboratories, #152 and #182 (both samples types) for which only data from the CGE tools were generated (Table M.5). All laboratories were able to identify in *E. coli* genome GMI17-004 the expected antimicrobial resistance genes, $bla_{\text{TEM-1B}}$, *mcr*-1, *sul*1, *sul*2, *tet*(A), *str*A, and *drf*A1 using own and CGE tools. Furthermore, the *aad*A1 gene was also expected in GMI17-004 and was observed using the CGE tool by all laboratories except for laboratory #185 using own tool. In addition, also the two genes, *aph*(3")-Ib and *aph*(6)-Id were expected in GMI17-004. These genes were only detected by using the CGE tool except for laboratory #185 which also reported the presence of these genes based on the use of their own tool. Laboratory #185 also reported *mdf*(A) using own tool. Interestingly, all laboratories reported the gene *str*B using own tools which was not expected in the reference genome. The *str*A gene was located on plasmid 2 and it might be that *str*B was not captured by the PacBio sequencing of the reference.

S. aureus

Almost all laboratories reported 'unknown MLST' as was also the expected result for the *S. aureus* genome GMI17-005 except for laboratories #141, #146, #170, and #185 using their own tool revealing an incorrect ST4307 (arcc-6, aroe-193, glpf-419, gmk-2, pta-7, tpi-58, yqil-52) or a novel ST (#146) (Table M.1).

Four of the laboratories provided data using own and CGE tools whereas the laboratories, #152 (both samples types), #182 (both samples types), and #185 GMI17-005-DNA for which only data from the CGE tools were generated (Table M.6). The *S. aureus* genome GMI17-005 solely contained the antimicrobial resistance gene, *mec*C, which was detected by all laboratories using the CGE tool and only by laboratories #182 and #185 using own tool. The other laboratories reported the *mec*A gene using their own tool.

The expected MLST *S. aureus* genome GMI17-006 was ST398 which was correctly reported by all laboratories using the CGE tool and their own tool with exception of laboratories #141 and #182 which reported ST4251 using own tool (both samples types) and ST2850 using CGE tool (BACT sample) (Table M.1).

All laboratories were able to identify in *S. aureus* genome GMI17-006 the expected antimicrobial resistance genes, *mec*A, *Isa*(B), *crf*, *Inu*(B), *fex*A, *tet*(M), and *drf*G using own and CGE tools (Table M.7). Furthermore, a few discrepancies were observed among the laboratories detecting the following expected antimicrobial resistance genes, *spc* (laboratory #185, own tool), *bla*Z (laboratory #156, both sample types using own tool, #182, DNA using CGE tool), and *tet*(K) (laboratory #156, and #182 using CGE tool on the BACT sample). In addition, the following laboratories, #141, #146, #156 and #170, all detected the gene, *nor*A, using own tools which was not expected in the reference genome. Similarly, laboratory #152 identified using the CGE tool the gene *cat*A1 present in the BACT sample. Interestingly, all laboratories reported the *str*-gene using both tools but were not expected in the reference genome.

Sequencing – Single Nucleotide Polymorphism

The raw reads of both the culture and corresponding DNA were mapped to the corresponding reference to identify SNP. No SNPs were observed in the genomes submitted by the laboratories with only one exception. One SNP was detected in relation to the GMI17-005-BACT sequence submitted from laboratory #170. This has been assessed as a spontaneous mutation and not due to a contamination.

Discussion

Laboratory #152 did not participate in the *Salmonella* trials. Laboratory #182 and #152 did not report any antimicrobial resistance genes for genome GMI17-001– GMI17-006 and GMI17-003–GMI17-006, respectively.

The majority of the submitted MLST data were correct and in line with the expected results using the CGE reference method. A few laboratories, #141, #146, #170 and #185 reported variants or closely related MLST using own tools compared to the reference indicating that these laboratories might consider the tools used to predict MLSTs.

Most of the submitted AMR genes were in concordance with the expected results. Some deviations, however, were observed especially related to laboratories reporting *str* genes using own tools and laboratories having problems to identify the following genes, *aac*(3)-lla in GMI17-002, *aph*(3")-lb and *aph*(6)-ld in GMI17-004, and *mec*A/*nor*A in GMI17-005 using own tools. Interestingly, laboratory #156 identified a completely different number of antimicrobial resistance genes compared to those expected in GMI17-003-BACT raw reads from #156 laboratory using CGE tools. It was evident that laboratory #156 submitted and analysed the wrong genome as being an outlier when assessing the coverage compared to the reference chromosome and the plasmids. This also included a high number of contigs and low N50.

One of the objectives for the ENGAGE PT was to assess a range of quality markers to evaluate the performance by the consortium partners. Overall, the 2017 PT showed that all laboratories performed satisfactorily, with minor exceptions. In general, laboratory #141 produced a high number of reads including a high percentage of mapping reads to the references, to plasmids and unmapped reads. This could potentially be explained by the laboratory running the test on a new large-scale NextSeq. In addition, also laboratory #185 revealed a high number of unmapped reads as well as a low percentage of reads mapping to the reference genome which also might be related to using a large-scale platform (HiSeq 2500). Laboratory #156 resulted an outlier for a few of the genomes in relation to the assembly conducted by the PT organizers, indicating a possible contamination of the genomes produced. It was not possible to identify the reason for the large assembly sizes but it apparently did not affect the results. It was also noteworthy that laboratory #170 obtained a Phred quality score (Q score) below Q30 (tentative QC threshold) for both *S. aureus* genomes indicating base call accuracy less than of 99.9 %, with the probability of an incorrect base call in more than 1 out of 1,000 bp. The laboratory has reported the use of an Illumina MiSeq platform which normally performs with a higher Q score.

Conclusions

The PT was a useful exercise as it allowed the ENGAGE consortium partners to assess the quality of their own data as well as to identify critical points for improvement. In general, all data were satisfactory but the PT organizer especially encourages laboratory #141 to optimize the sequencing procedures for the new NextSeq to avoid too many reads including unmapped reads to be produced, laboratory #156 to investigate where in the process genome GMI17-003-BACT was switched and also why some of the assemblies made by the PT organizer were too large and laboratory #185 to investigate the reason behind the low percentages of reads mapping to the reference genomes.

References

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM and Lund O, 2012. Multilocus sequence typing of total-genome-sequenced bacteria. Journal of Clinical Microbiology, 50(4):1355-1361

Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E and Hendriksen RS, 2015. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infectious Diseases, 15:174-0902.

Price MN, Dehal PS and Arkin AP, 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS.One, 5:e9490.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM and Larsen MV, 2012. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy, 67:2640-2644.

THE ENGAGE PROFICIENCY TEST REPORT 2017

TABLES AND FIGURES

Table M.1: Determined MLST for the bacterial culture and DNA received

Reference strain MLST:			GMI17-001	GMI17-002	GMI17-003	GMI17-004	GMI17-005	GMI17-006
			ST-142	**ST-14**	**ST-448**	**ST-10**	**Unknown ST**	**ST-398**
BACT	#141	Own tool	ST-377	ST-14	ST-448	ST-10	ST-4307	ST-4251
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
DNA		Own tool	ST-377	ST-14	ST-448	ST-10	ST-4307	ST-4251
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
BACT	#146	Own tool	ST-142	ST-14	ST-448	Novel type	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
DNA		Own tool	ST-142	ST-14	ST-448	Novel type	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
BACT	#152	Own tool	-	-		ST-10	Unknown ST	ST-398
		CGE tool	-	-	ST-448	ST-10	Unknown	ST-398
DNA		Own tool	-	-		ST-10	Unknown ST	ST-398
		CGE tool	-	-	ST-448	ST-10	Unknown	ST-398
BACT	#156	Own tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
		CGE tool	ST-142	ST-14	Unknown ST	ST-10	Unknown	ST-398
DNA		Own tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
BACT	#170	Own tool	ST-142	ST-14	ST-448	Unknown	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
DNA		Own tool	ST-142	ST-14	ST-448	Unknown	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
BACT	#182	Own tool	-	-	-	-	-	-
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-2850
DNA		Own tool	-	-	-	-	-	-
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
BACT	#185	Own tool	ST-142	ST-14	ST-448	Novel ST	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398
DNA		Own tool	ST-142	ST-14	ST-448	Novel ST	ST-4307	ST-398
		CGE tool	ST-142	ST-14	ST-448	ST-10	Unknown	ST-398

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.2: Determined antimicrobial resistance genes for the bacterial culture and DNA received

Reference antimicrobial resistance genes			GMI17-001									
			-	-	*bla*$_{TEM-1B}$	-	*sul2*	*tet*(A)	*drf*A14	-	*aph*(6)-Id	*aph*(3")-Ib
BACT	#141	Own tool	*str*A (partial)	*str*B	TEM-1B	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool	*str*A (partial)	*str*B	TEM-1B	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
BACT	#146	Own tool	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
BACT	#152	Own tool	-	-	-	-	-	-	-	-	-	-
		CGE tool	-	-	-	-	-	-	-	-	-	-
DNA		Own tool	-	-	-	-	-	-	-	-	-	-
		CGE tool	-	-	-	-	-	-	-	-	-	-
BACT	#156	Own tool	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
BACT	#170	Own tool		*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool		*str*B	*bla*$_{TEM-1B}$	mcr-1	*sul2*	*tet*(A)	*drf*A14			
		CGE tool			*bla*$_{TEM-1B}$	mcr-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool	-	-	-	-	-	-	-	-	-	-
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
BACT	#185	Own tool	*str*A	*str*B	TEM-1	*mcr*-1	*sul2*	*tet*(A)-1	*drf*A14	*aac*(6')-Iy	*aph*(6)-Id	
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib
DNA		Own tool	*str*A	*str*B	TEM-1	*mcr*-1	*sul2*	*tet*(A)-1	*drf*A14	*aac*(6')-Iy	*aph*(6)-Id	
		CGE tool			*bla*$_{TEM-1B}$	*mcr*-1	*sul2*	*tet*(A)	*drf*A14		*aph*(6)-Id	*aph*(3")-Ib

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.3: Determined antimicrobial resistance genes for the bacterial culture and DNA received

Reference antimicrobial resistance genes			GMI17-002											
			***aad*A1**	**-**	**-**		***aac*(3)-IIa**	**-**	**-**	***aac*(6')-Ib-cr**	***aac*(6')-Ib**	***bla*$_{TEM-1A}$**	**-**	**-**
BACT	#141	Own tool		*aad*A24 (partial)			*aac*(3)-IIa	*aac*(6')Ib-cr (partial)		*aac*(6')Ib-cr (partial)		*bla*$_{TEM-1A}$		
		CGE tool					*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
DNA		Own tool		*aad*A24 (partial)				*aac*(6')Ib-cr (partial)		*aac*(6')Ib-cr (partial)		*bla*$_{TEM-1A}$		
		CGE tool					*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
BACT	#146	Own tool	*aad*A1		*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$	*bla*$_{TEM-1B}$	
		CGE tool	*aad*A1		*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr	*aac*(6')-Ib	*aac*(6')Ib-cr	*aac*(6')-Ib		*bla*$_{TEM-1B}$	
DNA		Own tool	*aad*A1		*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$	*bla*$_{TEM-1B}$	
		CGE tool	*aad*A1		*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr	*aac*(6')-Ib	*aac*(6')Ib-cr	*aac*(6')-Ib		*bla*$_{TEM-1B}$	
BACT	#156	Own tool		*aad*A24	*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
		CGE tool			*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
DNA		Own tool		*aad*A24	*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
		CGE tool						*aac*(6')Ib-cr	*aac*(6')-Ib	*aac*(6')Ib-cr	*aac*(6')-Ib	*bla*$_{TEM-1A}$		
BACT	#170	Own tool			*aac*(3)-IId			*aac*(6')-Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
		CGE tool			*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
DNA		Own tool			*aac*(3)-IId			*aac*(6')Ib-cr-1		*aac*(6')Ib-cr-1			*bla*$_{TEM-1B}$	
		CGE tool			*aac*(3)-IId		*aac*(3)-IIa	*aac*(6')Ib-cr		*aac*(6')Ib-cr			*bla*$_{TEM-1B}$	
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool			*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr		*bla*$_{TEM-1A}$		
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool			*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr			*bla*$_{TEM-1B}$	
BACT	#185	Own tool				*aac*(6')-Iy	*aac*(3)-IIa	*aac*(6')-Ib-cr		*aac*(6')-Ib-cr				TEM-191-p
		CGE tool			*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr			*bla*$_{TEM-1B}$	
DNA		Own tool				*aac*(6')-Iy	*aac*(3)-IIa	*aac*(6')-Ib-cr		*aac*(6')-Ib-cr				TEM-191-p
		CGE tool			*aac*(3)-IId			*aac*(6')Ib-cr		*aac*(6')Ib-cr			*bla*$_{TEM-1B}$	

Reference antimicrobial resistance genes			GMI17-002, continued												
			***bla*$_{CTX-M-15}$**	**-**	***bla*$_{OXA-1}$**	***bla*$_{OXA-9}$**	***bla*$_{OXA-10}$**	***mph*(A)**	**-**	***cat*B4**	***cml*A1**	**ARR-3**	***sul*1**	**-**	**-**	**-**
BACT	#141	Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$(partial)		
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$(partial)		
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
BACT	#146	Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1			
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1			
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
BACT	#156	Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)	*cat*B3		*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)	*cat*B3		*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	sul1	*bla*$_{DHA-1}$		
BACT	#170	Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1			
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1			
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
BACT	#185	Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)			*cml*A1	ARR-3	*sul*1		*gyr*A_SET [83:S-Y;87:D-G]	*par*C_SET [57:T-S;80:S-I]
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		
DNA		Own tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)			*cml*A1	ARR-3	*sul*1		*gyr*A_SET [83:S-Y;87:D-G]	*par*C_SET [57:T-S;80:S-I]
		CGE tool	*bla*$_{CTX-M-15}$	*bla*$_{NDM-1}$	*bla*$_{OXA-1}$	*bla*$_{OXA-9}$	*bla*$_{OXA-10}$	*mph*(A)		*cat*B4	*cml*A1	ARR-3	*sul*1	*bla*$_{DHA-1}$		

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.4: Determined antimicrobial resistance genes for the bacterial culture and DNA received

Reference antimicrobial resistance genes			GMI17-003												
			***aad*A5**	***aac*(3)-IId**	***aac*(6')Ib-cr**	**bla_{TEM-1B}**	**bla_{NDM-7}**	**bla_{OXA-1}**	**-**	***mph*(A)**	***cat*A1**	***cat*B3**	***sul*1**	***dfr*A17**	**-**	**-**
BACT	#141	Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#146	Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#152	Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#156	Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
		CGE tool				bla_{TEM-1B}					*cat*A1					
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#170	Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3		*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3		*dfr*A17		
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
BACT	#185	Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}	*mdf*(A)	*mph*(A)	*cat*A1	*cat*B3		*dfr*A17	*gyr*A_EC2[83:S-L;87:D-N]	*par*C_EC2[80:S-I;84:E-G;88:L-Q]
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		
DNA		Own tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}	*mdf*(A)	*mph*(A)	*cat*A1	*cat*B3		*dfr*A17	*gyr*A_EC2[83:S-L;87:D-N]	*par*C_EC2[80:S-I;84:E-G;88:L-Q]
		CGE tool	*aad*A5	*aac*(3)-IId	*aac*(6')Ib-cr	bla_{TEM-1B}	bla_{NDM-7}	bla_{OXA-1}		*mph*(A)	*cat*A1	*cat*B3	*sul*1	*dfr*A17		

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.5: Determined antimicrobial resistance genes for the bacterial culture and DNA received

			GMI17-004											
Reference antimicrobial resistance genes			***aad*A1**	***str*A**	**-**	***bla*$_{TEM-1B}$**	***mcr*-1**	***sul*1**	***sul*2**	***tet*(A)**	***drf*A1**	***aph*(3")-Ib**	***aph*(6)-Id**	**-**
BACT	#141	Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#146	Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#152	Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#156	Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#170	Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1_1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	*aad*A1	*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1_1	*sul*1	*sul*2	*tet*(A)	*drf*A1			
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
BACT	#185	Own tool		*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1		*aph*(6)-Id	*mdf*(A)
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	
DNA		Own tool		*str*A	*str*B	*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1		*aph*(6)-Id	*mdf*(A)
		CGE tool	*aad*A1			*bla*$_{TEM-1B}$	*mcr*-1	*sul*1	*sul*2	*tet*(A)	*drf*A1	*aph*(3")-Ib	*aph*(6)-Id	

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.6: Determined antimicrobial resistance genes for the bacterial culture and DNA received

Reference antimicrobial resistance genes			GMI17-005		
			-	***mec*C**	-
BACT	#141	Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
DNA		Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
BACT	#146	Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
DNA		Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
BACT	#152	Own tool	-	-	-
		CGE tool		*mec*C	
DNA		Own tool	-	-	-
		CGE tool		*mec*C	
BACT	#156	Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
DNA		Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
BACT	#170	Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
DNA		Own tool	*mec*A		*nor*A
		CGE tool		*mec*C	
BACT	#182	Own tool	-	-	-
		CGE tool		*mec*C	
DNA		Own tool	-	-	-
		CGE tool		*mec*C	
BACT	#185	Own tool		*mec*C	
		CGE tool		*mec*C	
DNA		Own tool		*mec*C	
		CGE tool	-	-	-

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Table M.7: Determined antimicrobial resistance genes for the bacterial culture and DNA received

Reference antimicrobial resistance genes			GMI17-006												
				spc	***blaZ***	***mecA***	**-**	***Isa*(B)**	***cfr***	***Inu*(B)**	***fexA***	***tet*(K)**	***tet*(M)**	***drfG***	**-**
BACT	#141	Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
DNA		Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#146	Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
DNA		Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#152	Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	*catA1*
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#156	Own tool	*str*	*spc*		*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*		*tet*(M)	*dfrG*	
DNA		Own tool	*str*	*spc*		*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#170	Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
DNA		Own tool	*str*	*spc*	*blaZ*	*mecA*	*norA*	*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#182	Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*str*	*spc*		*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*		*tet*(M)	*dfrG*	
DNA		Own tool	-	-	-	-	-	-	-	-	-	-	-	-	-
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
BACT	#185	Own tool	*str*		*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
DNA		Own tool	*str*		*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	
		CGE tool	*str*	*spc*	*blaZ*	*mecA*		*Isa*(B)	*cfr*	*Inu*(B)	*fexA*	*tet*(K)	*tet*(M)	*dfrG*	

Data for the CGE tool were provided by PT-organizer and marked in light grey.

Legend to the following box plot figures (Figures M.1 to M.29):

Plot A shows results for all the available samples (excluding the ones with wrong MLST sequence type).

Red and blue lines indicate ± 2 and ± 3 standard deviations, respectively.

Plot B is an extract from plot A and shows the interquartile ranges of the distribution from plot A.

Boxplot points (red dots) indicate outlying values.

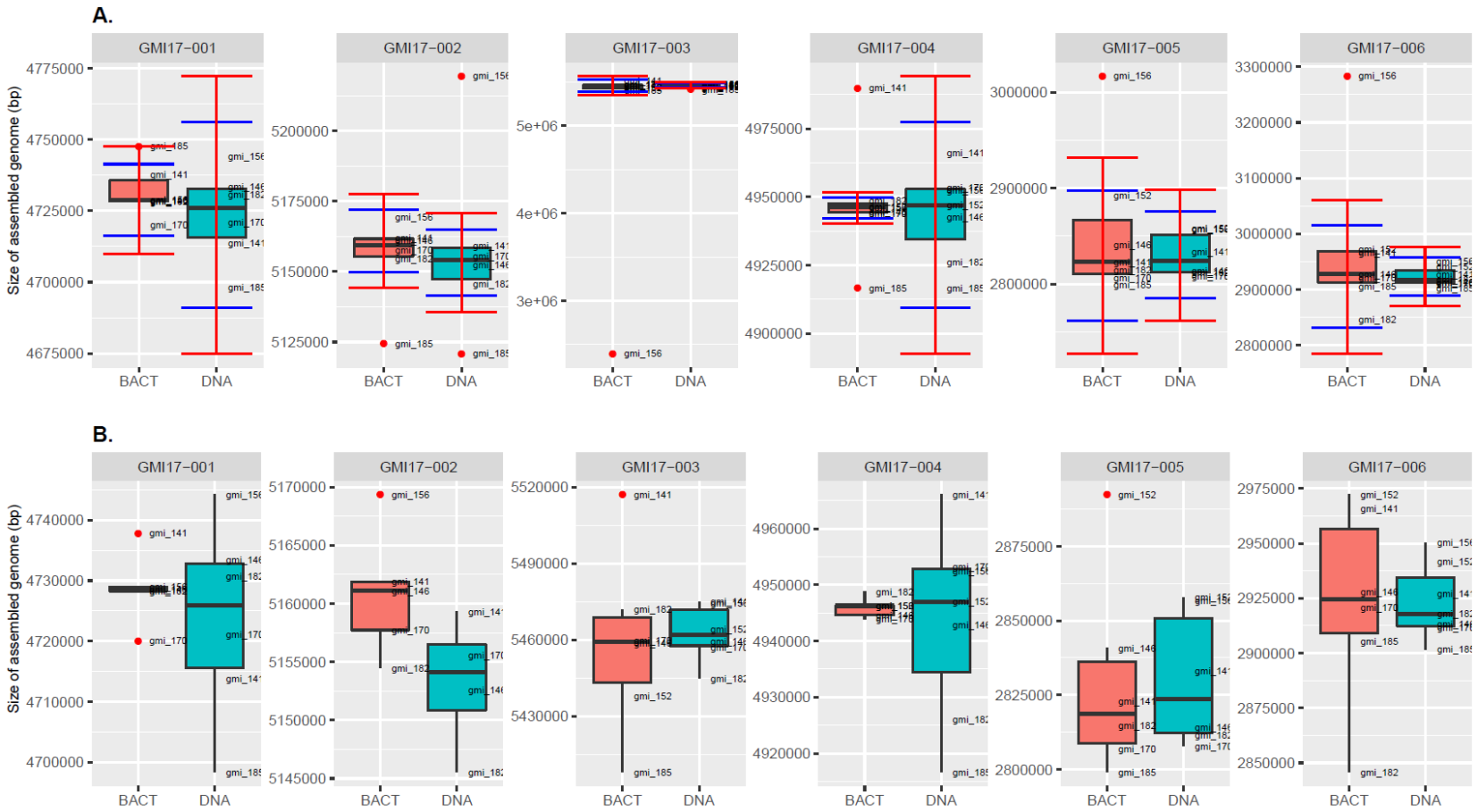


Figure M.1: Size of assembled genome

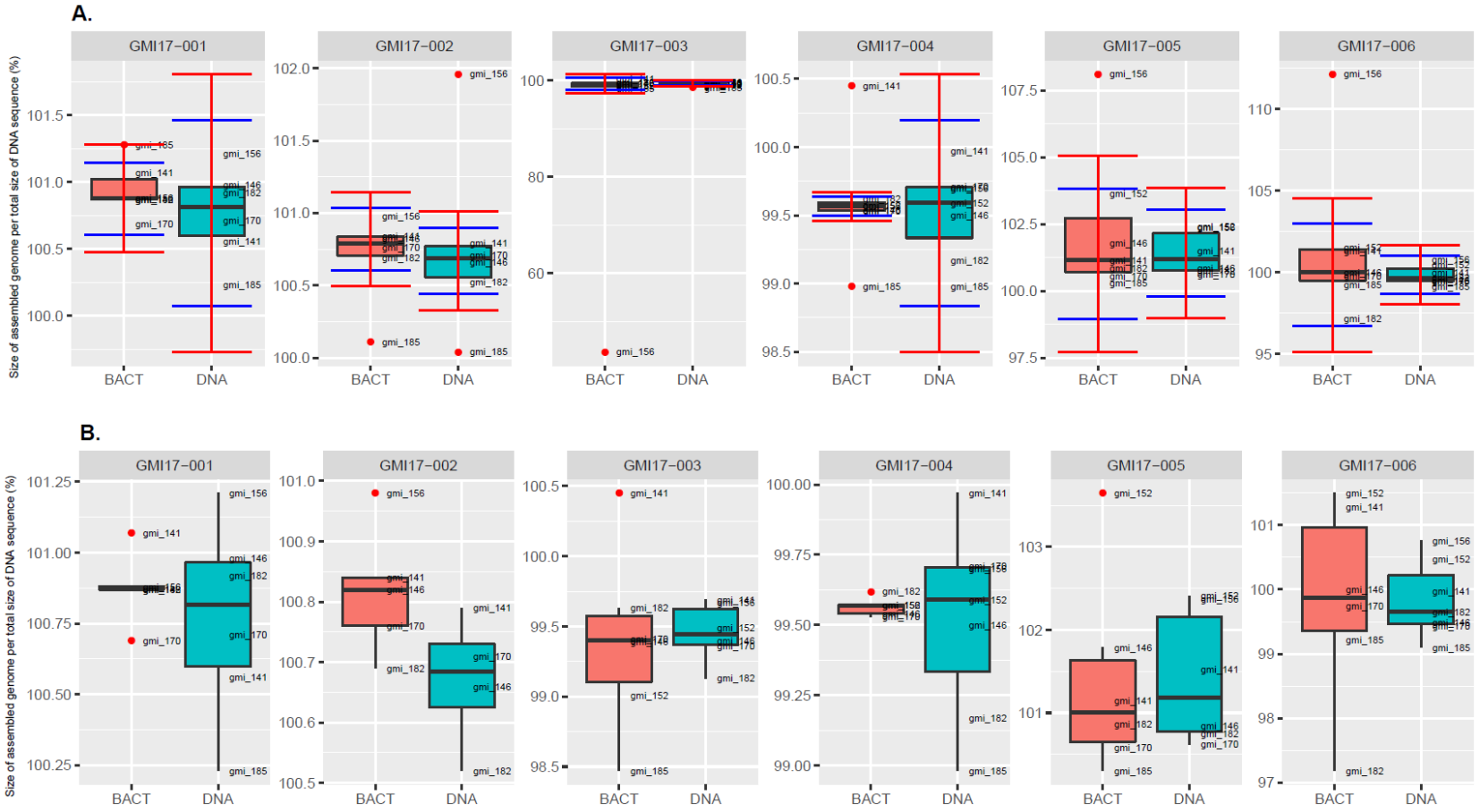


Figure M.2: Size of assembled genome per total size of DNA sequence

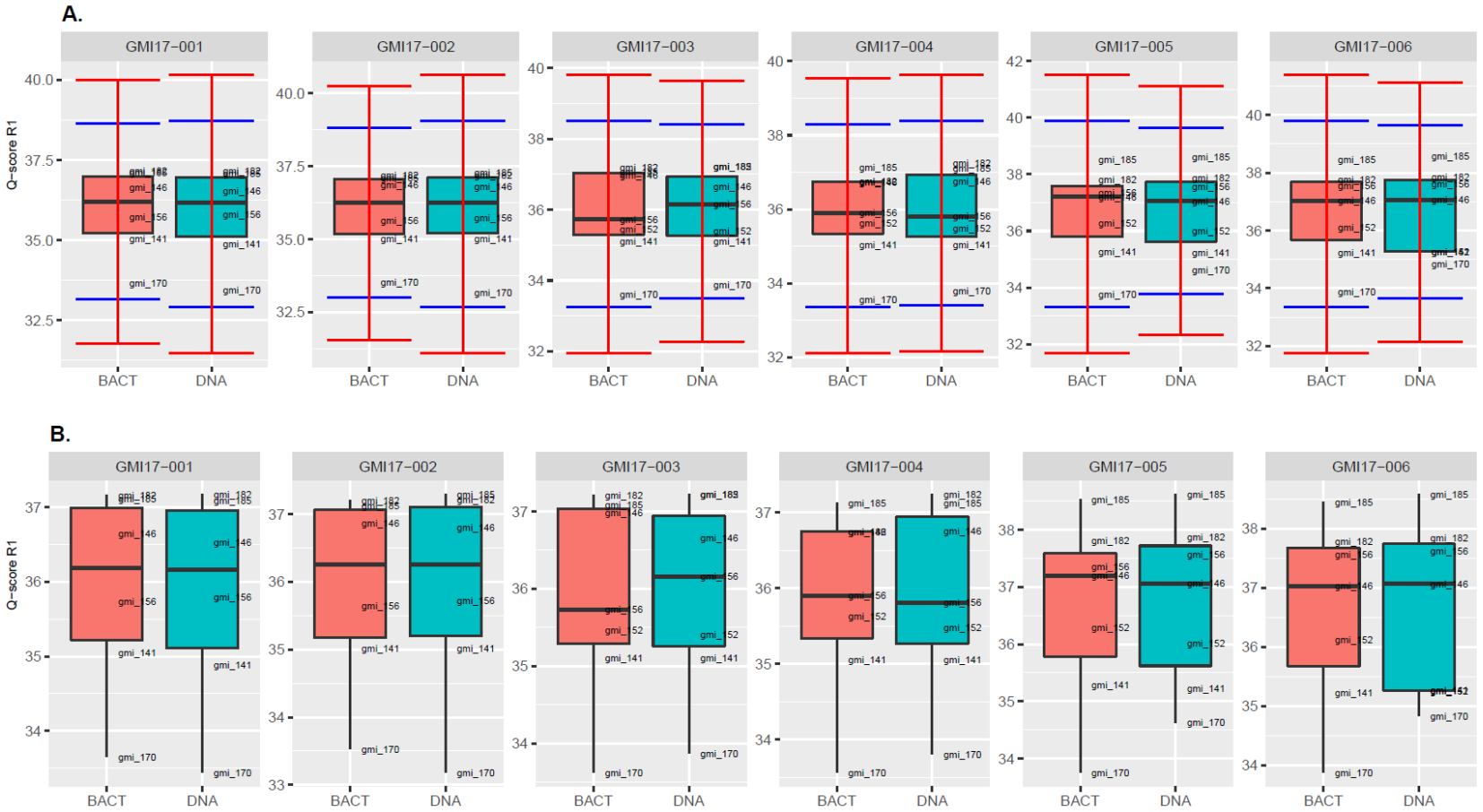


Figure M.3: Q-score, R1

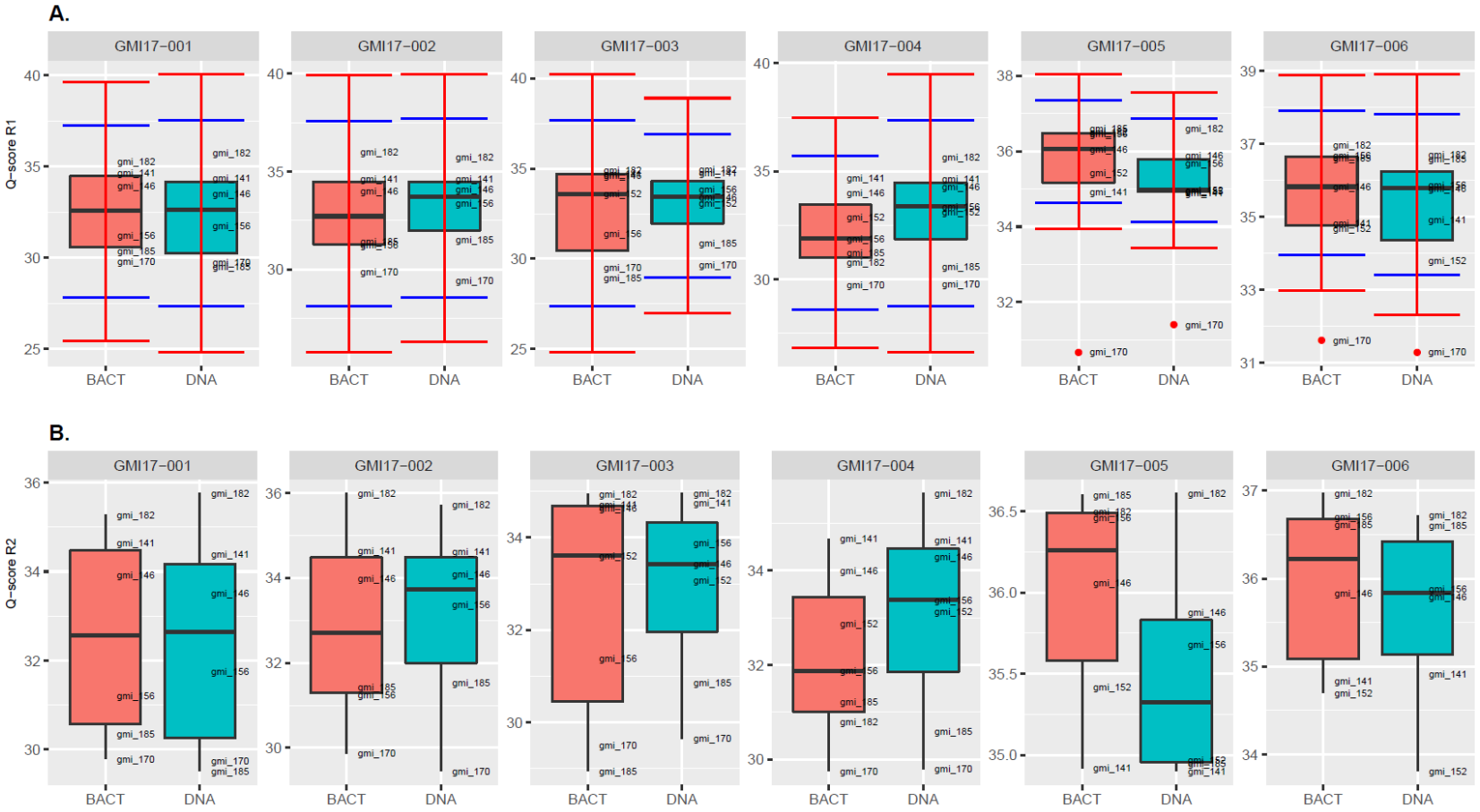


Figure M.4: Q-score, R2

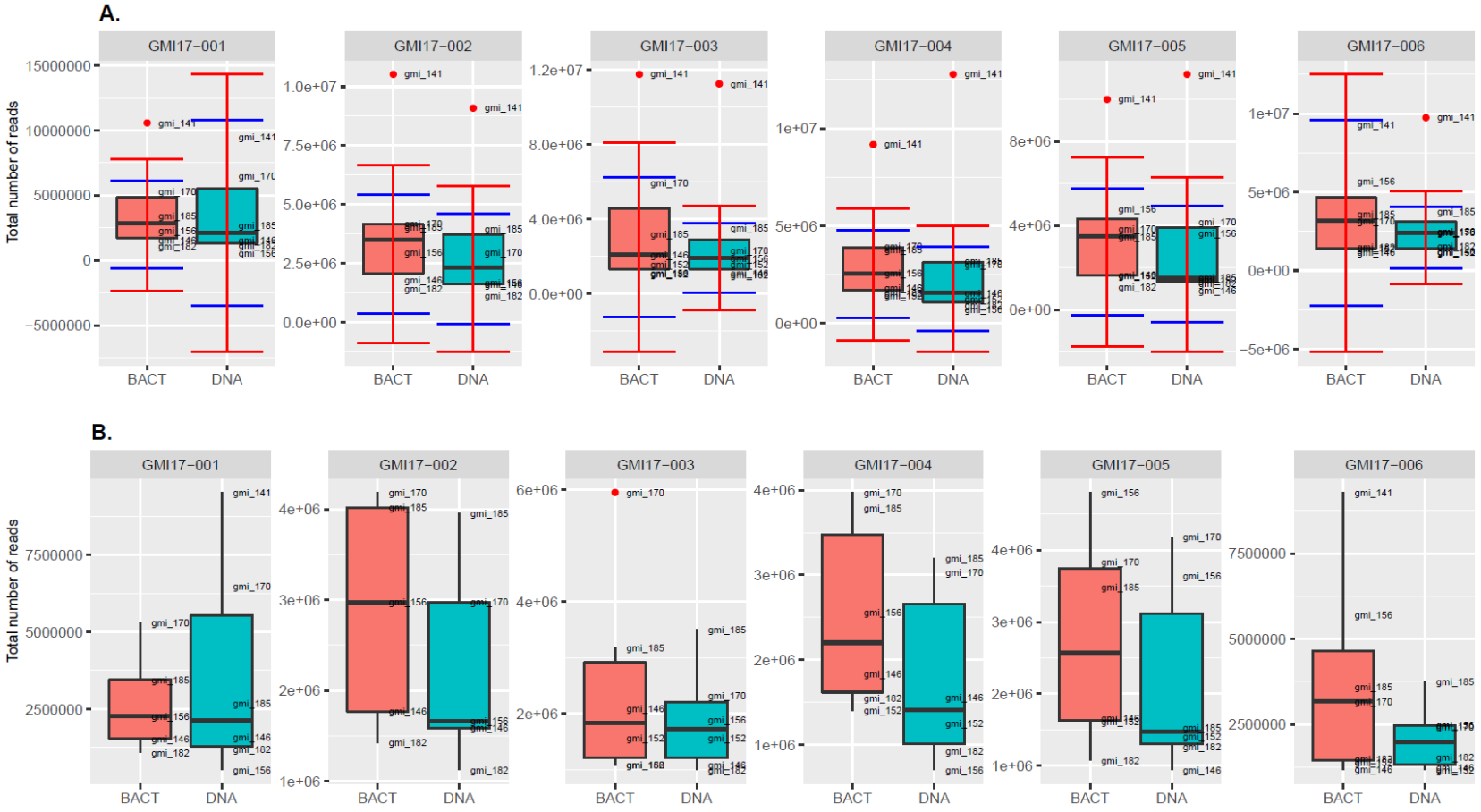


Figure M.5: Total number of reads

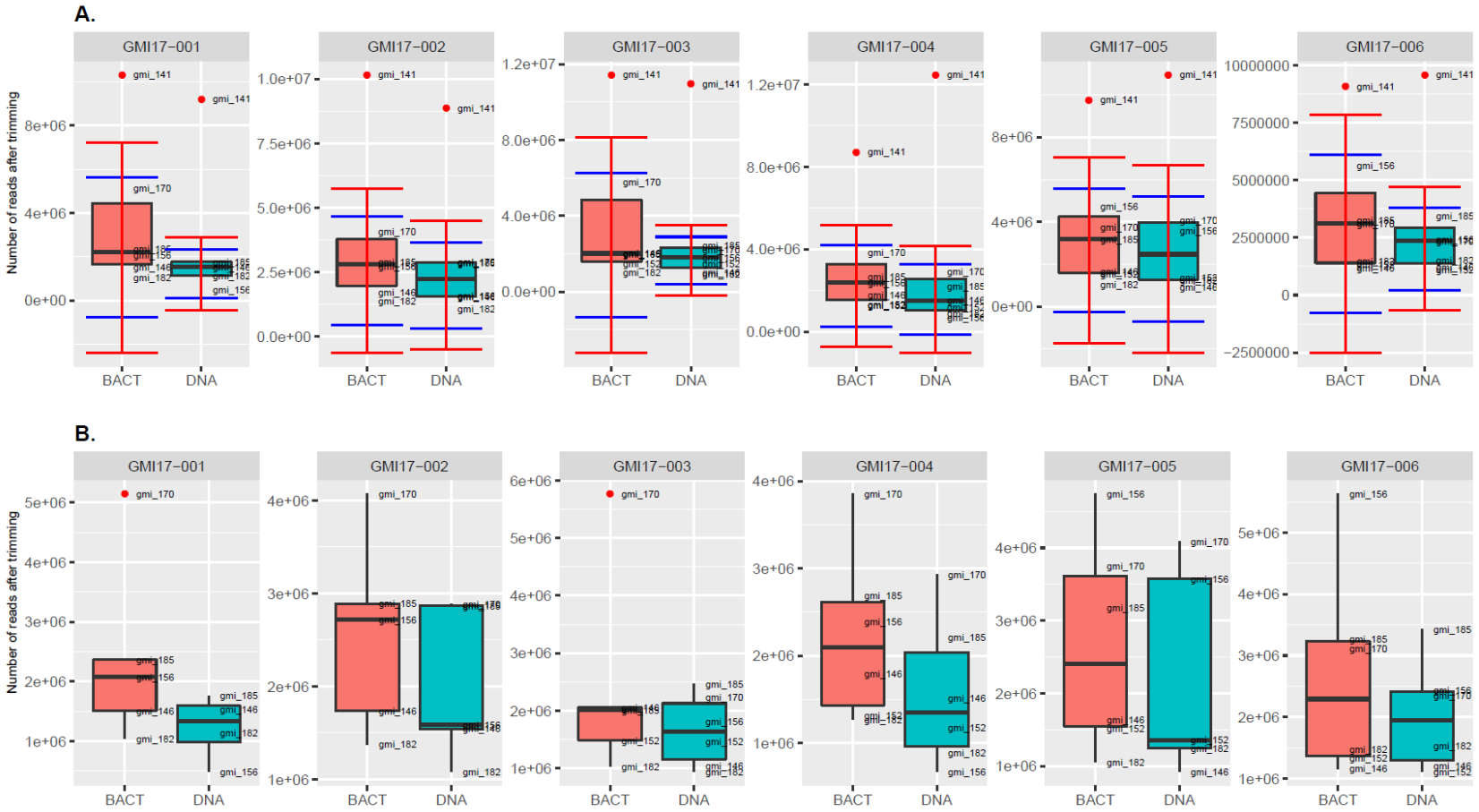


Figure M.6: Number of reads after trimming

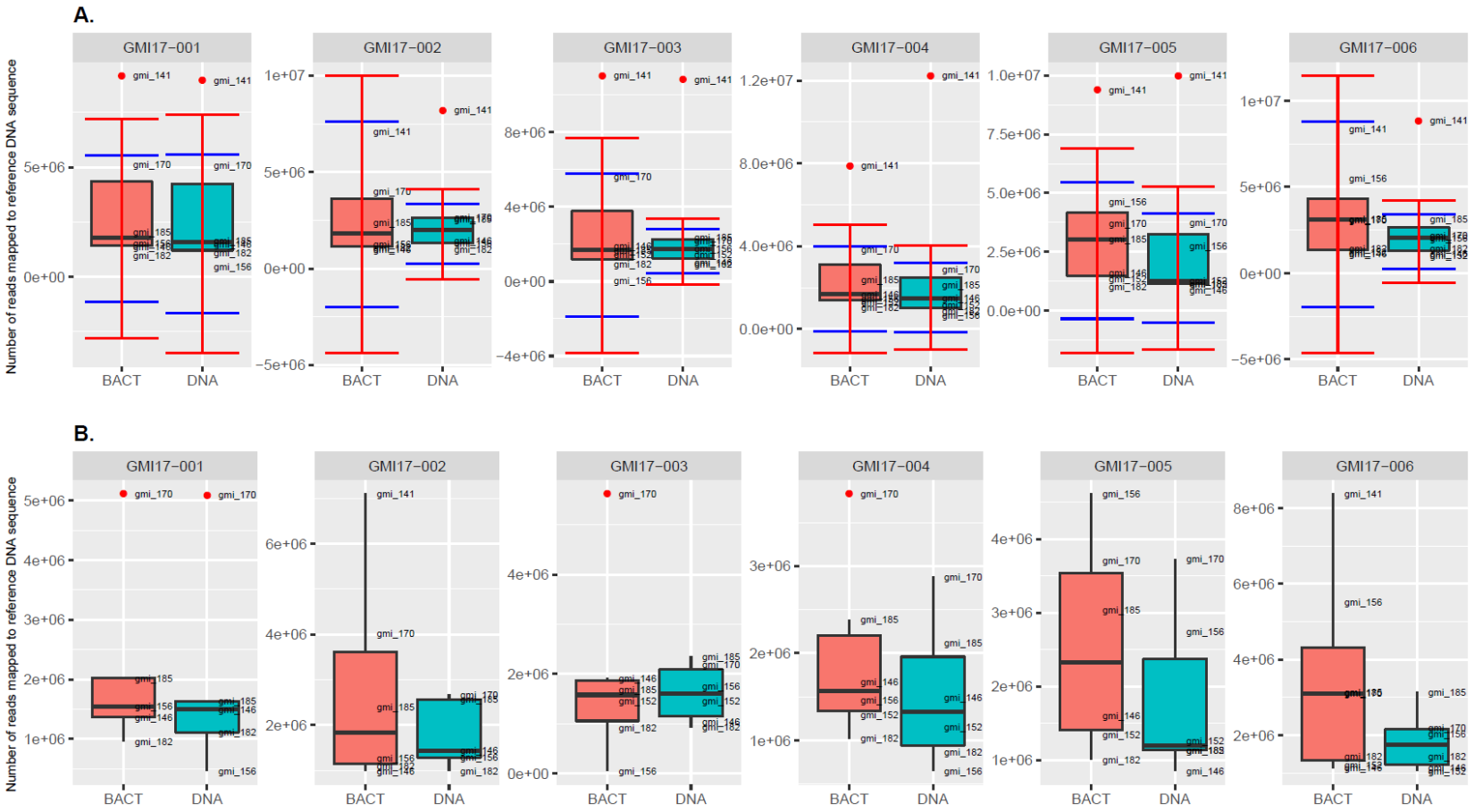


Figure M.7: Number of reads mapped to reference DNA sequence

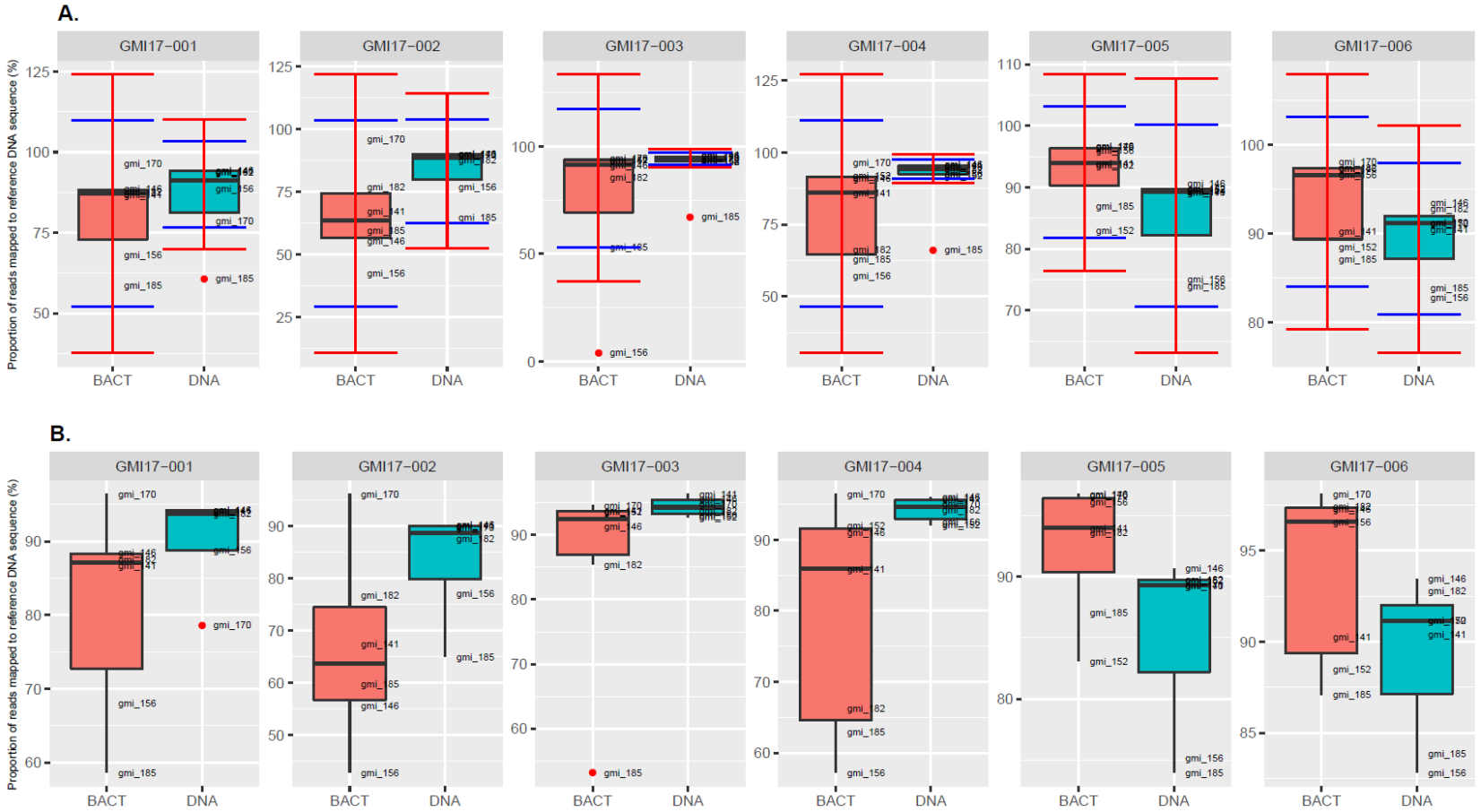


Figure M.8: Proportion of reads mapped to reference DNA sequence

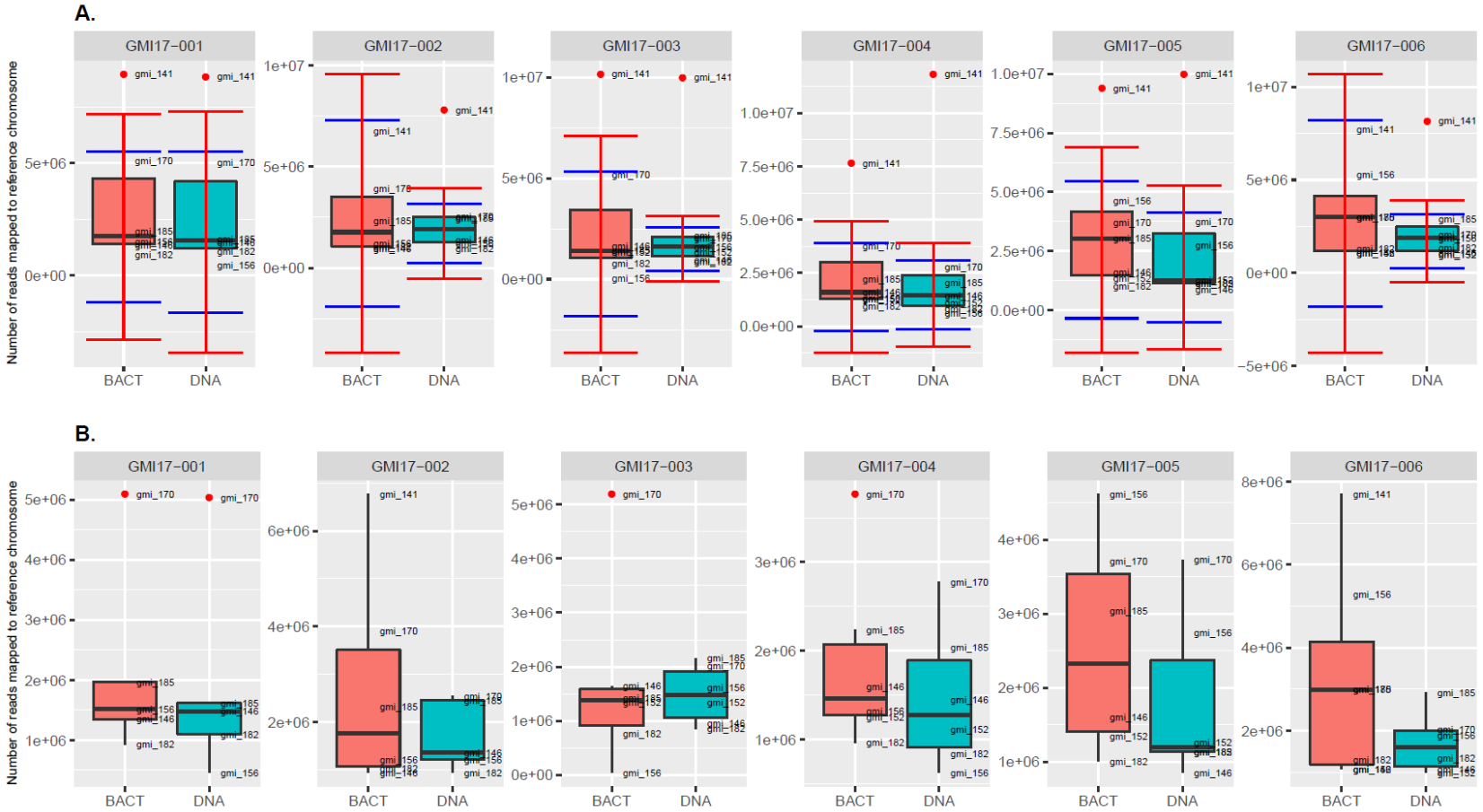


Figure M.9: Number of reads mapped to reference chromosome

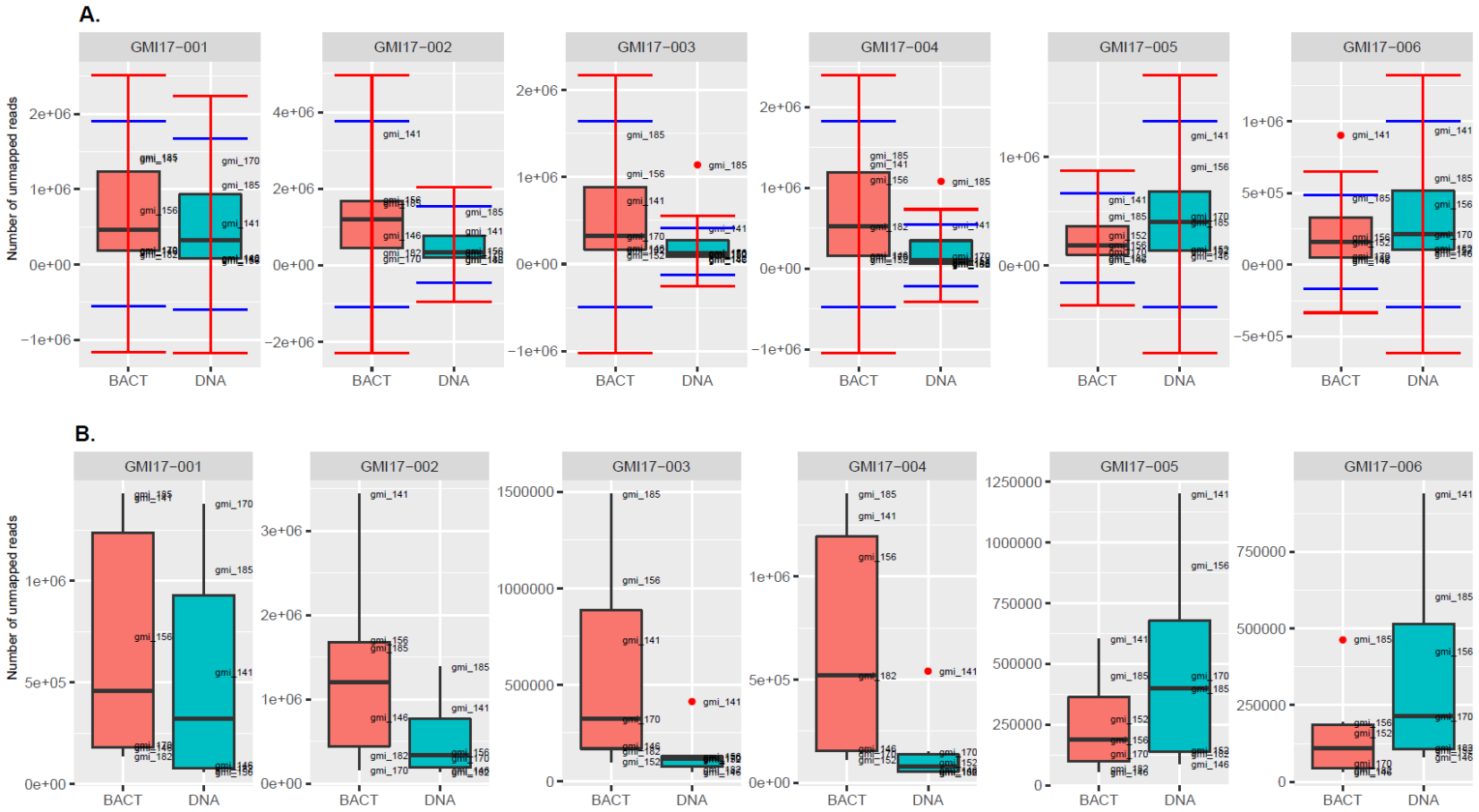


Figure M.10: Number of unmapped reads

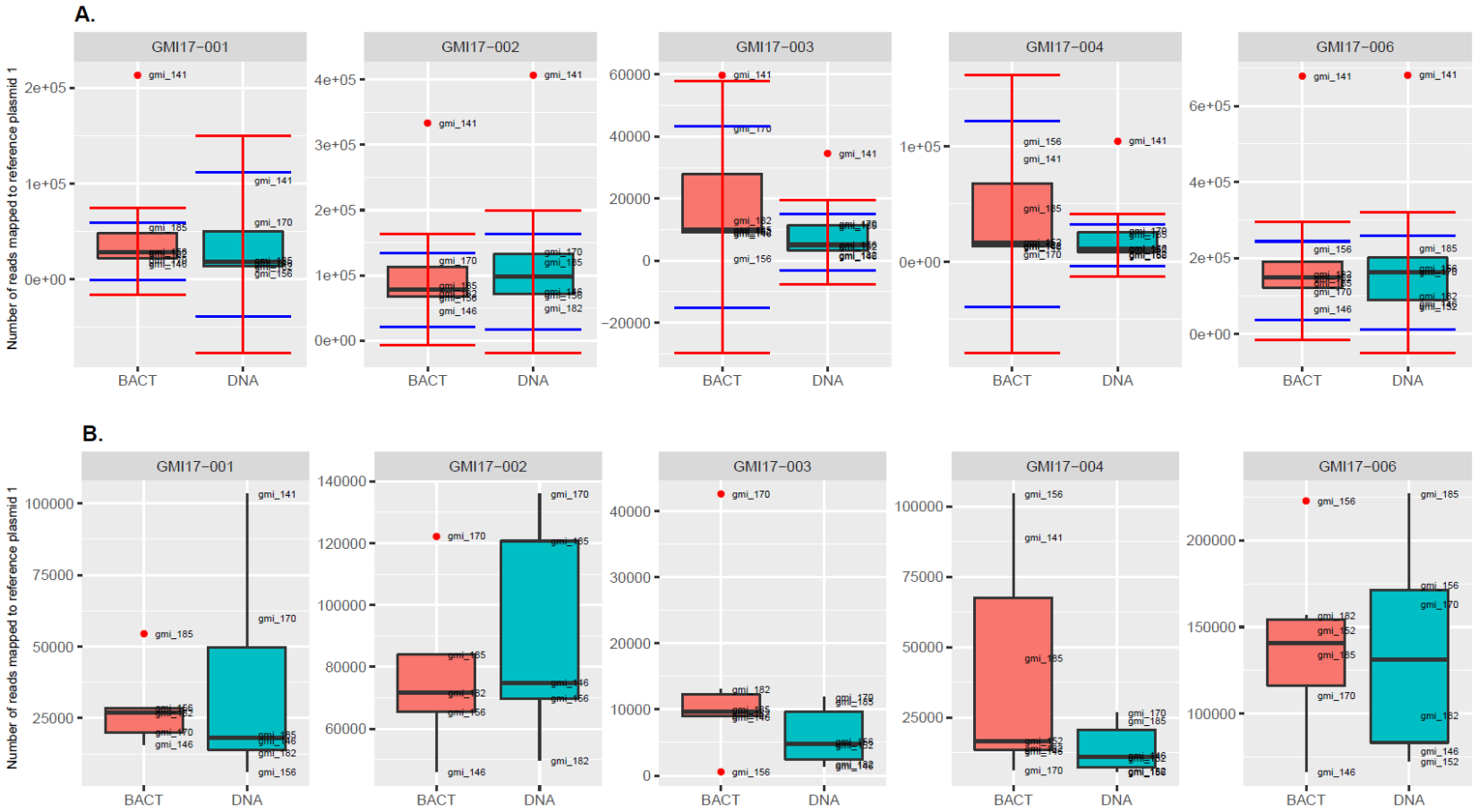


Figure M.11: Number of reads mapped to reference plasmid 1

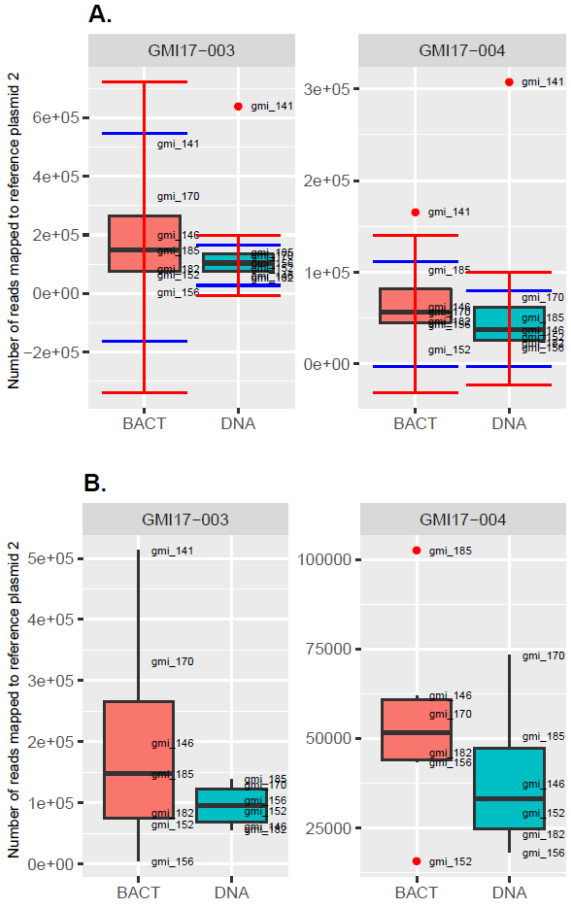


Figure M.12: Number of reads mapped to reference plasmid 2

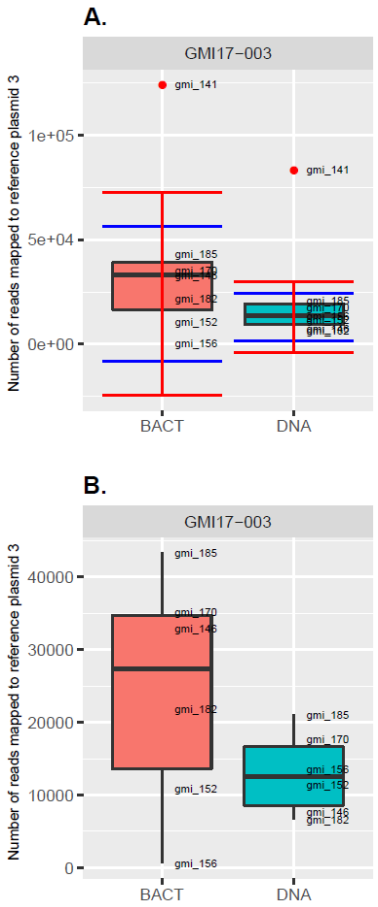


Figure M.13: Number of reads mapped to reference plasmid 3

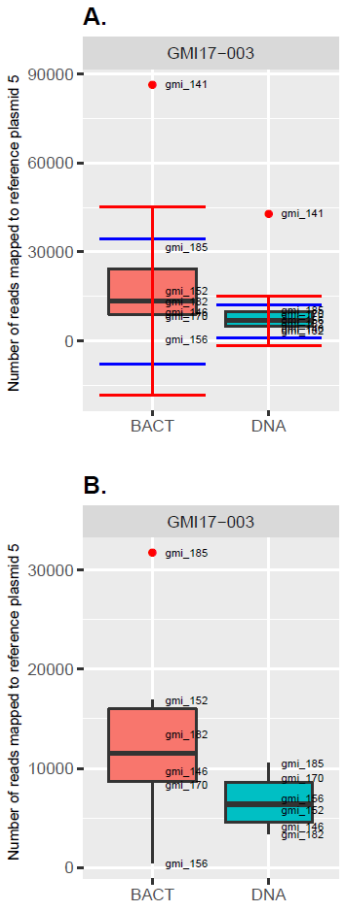


Figure M.14: Number of reads mapped to reference plasmid 4

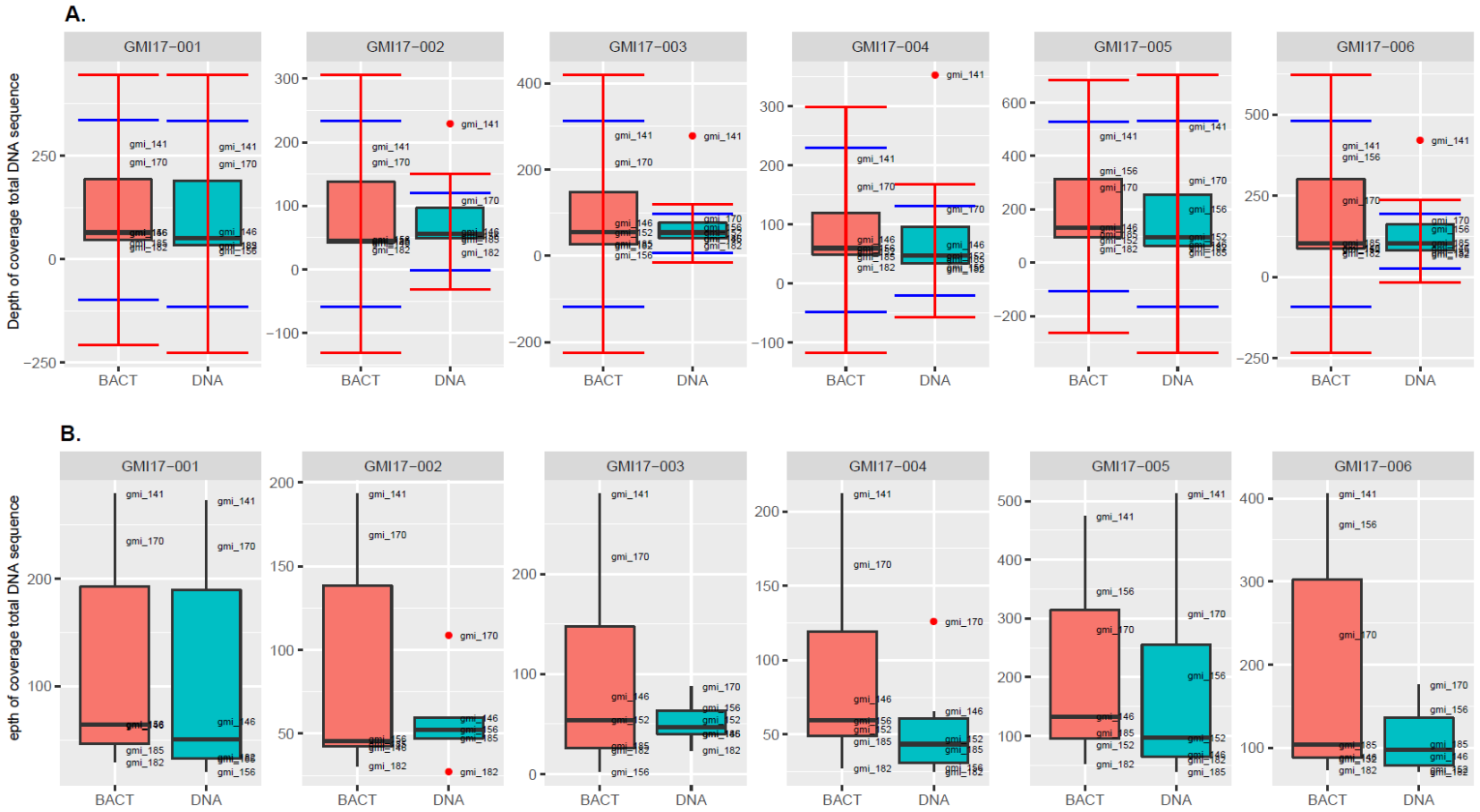


Figure M.15: Depth of coverage, total DNA sequence

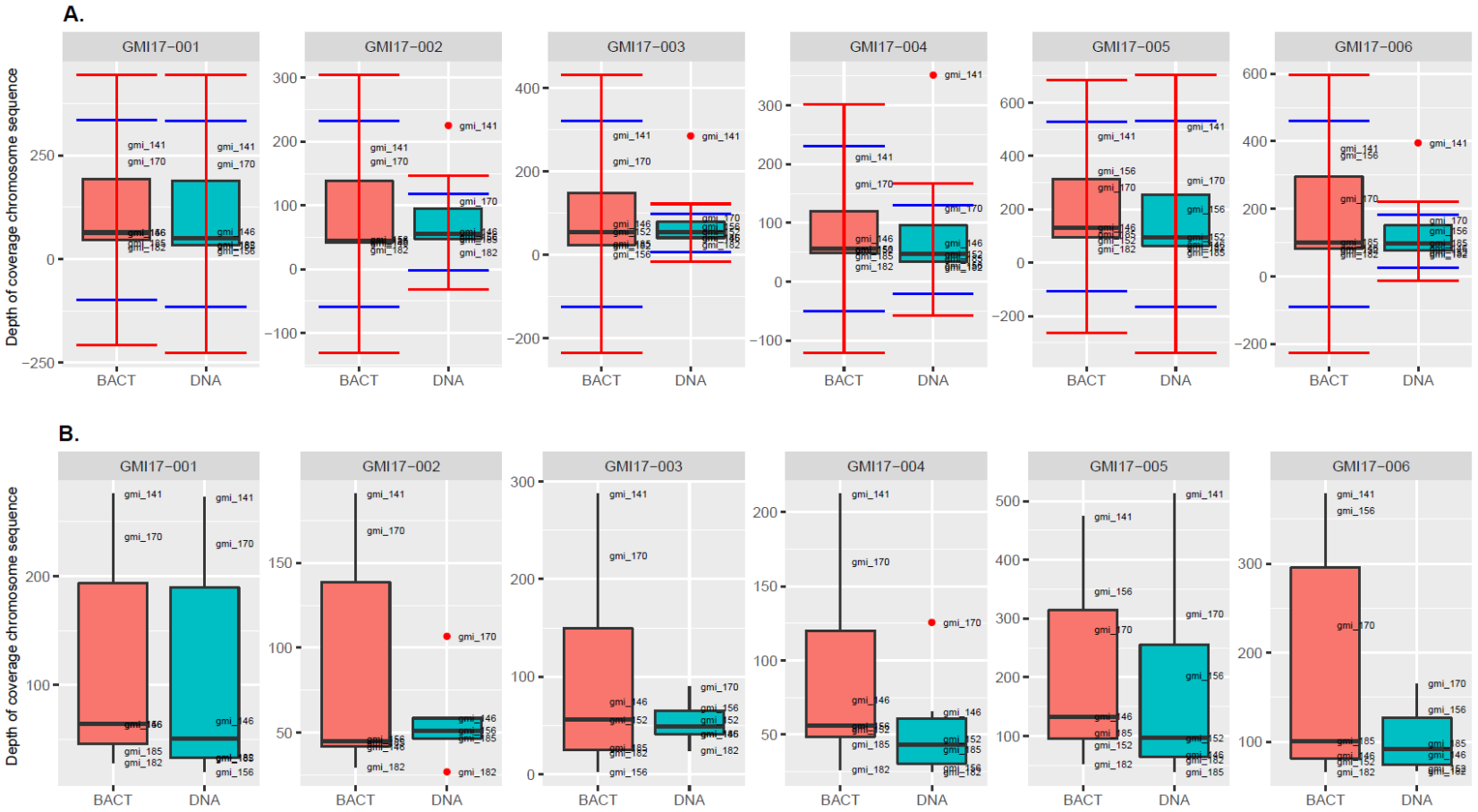


Figure M.16: Depth of coverage chromosome sequence

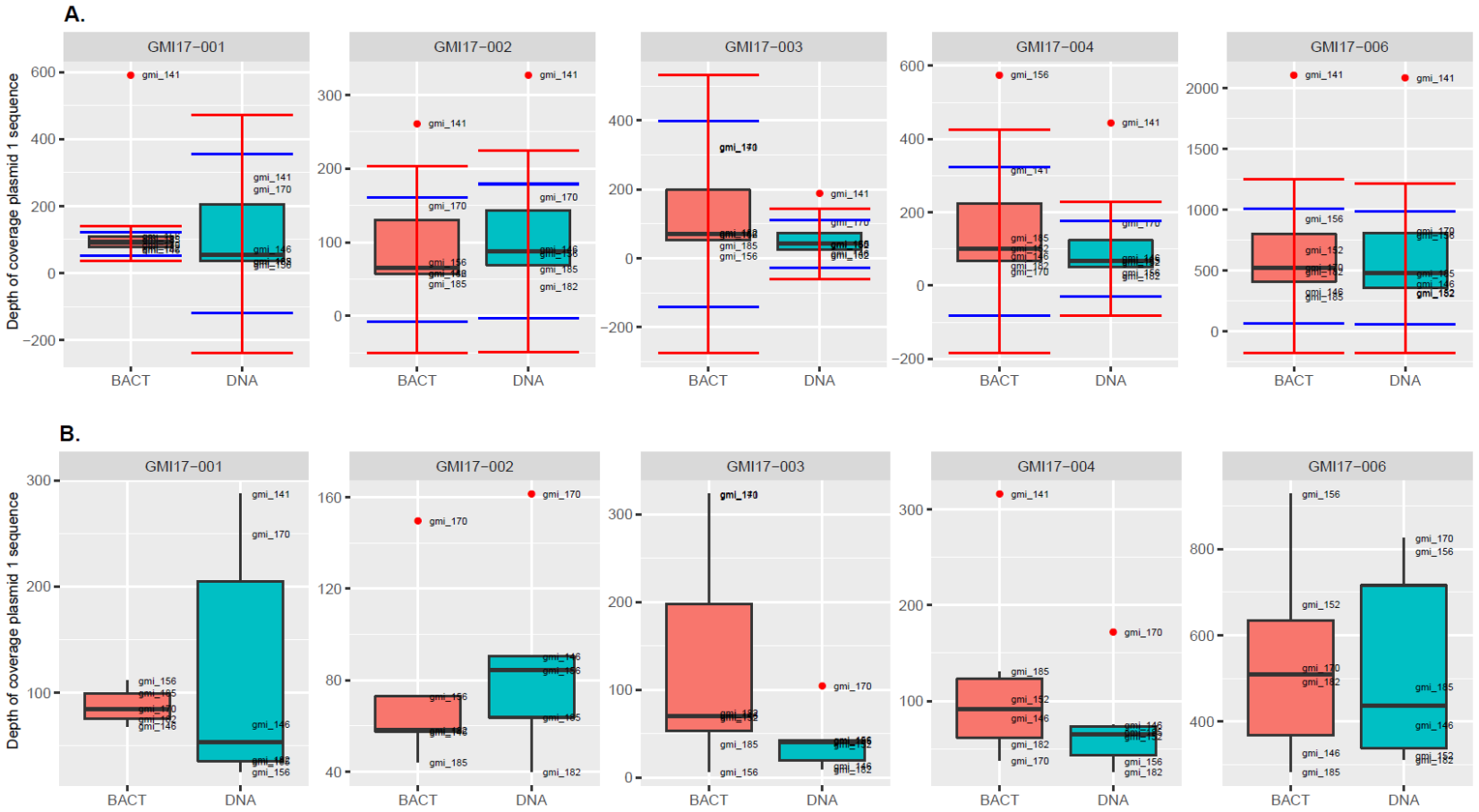


Figure M.17: Depth of coverage, plasmid 1 sequence

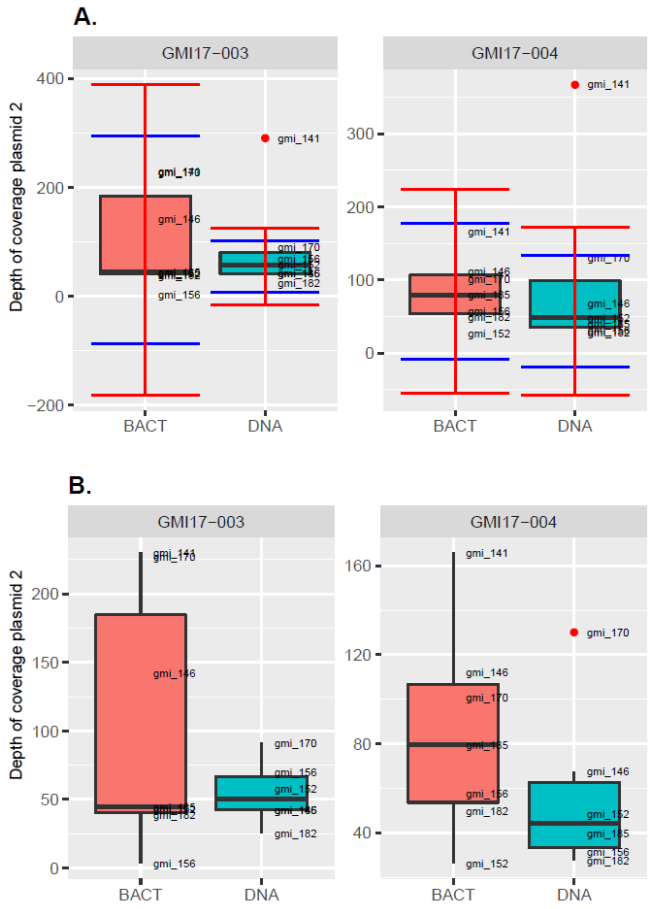


Figure M.18: Depth of coverage, plasmid 2 sequence

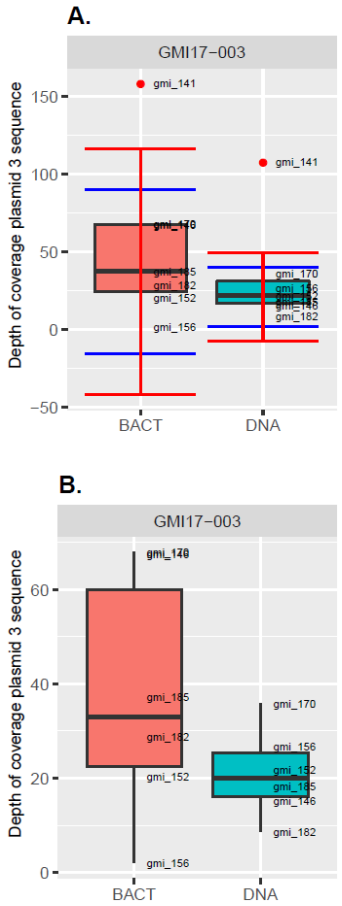


Figure M.19: Depth of coverage, plasmid 3 sequence

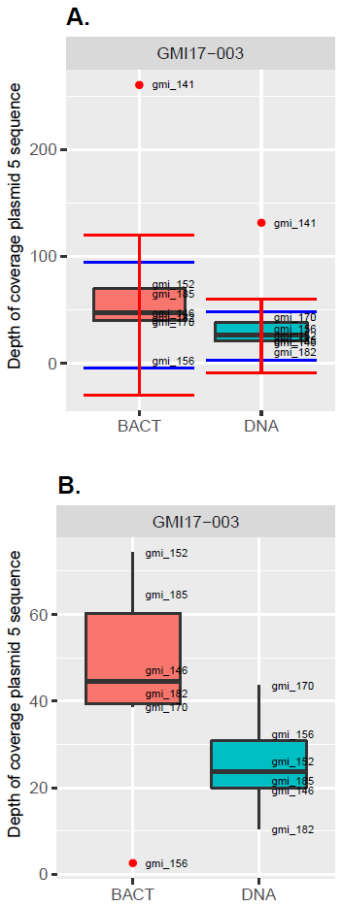


Figure M.20: Depth of coverage, plasmid 4 sequence

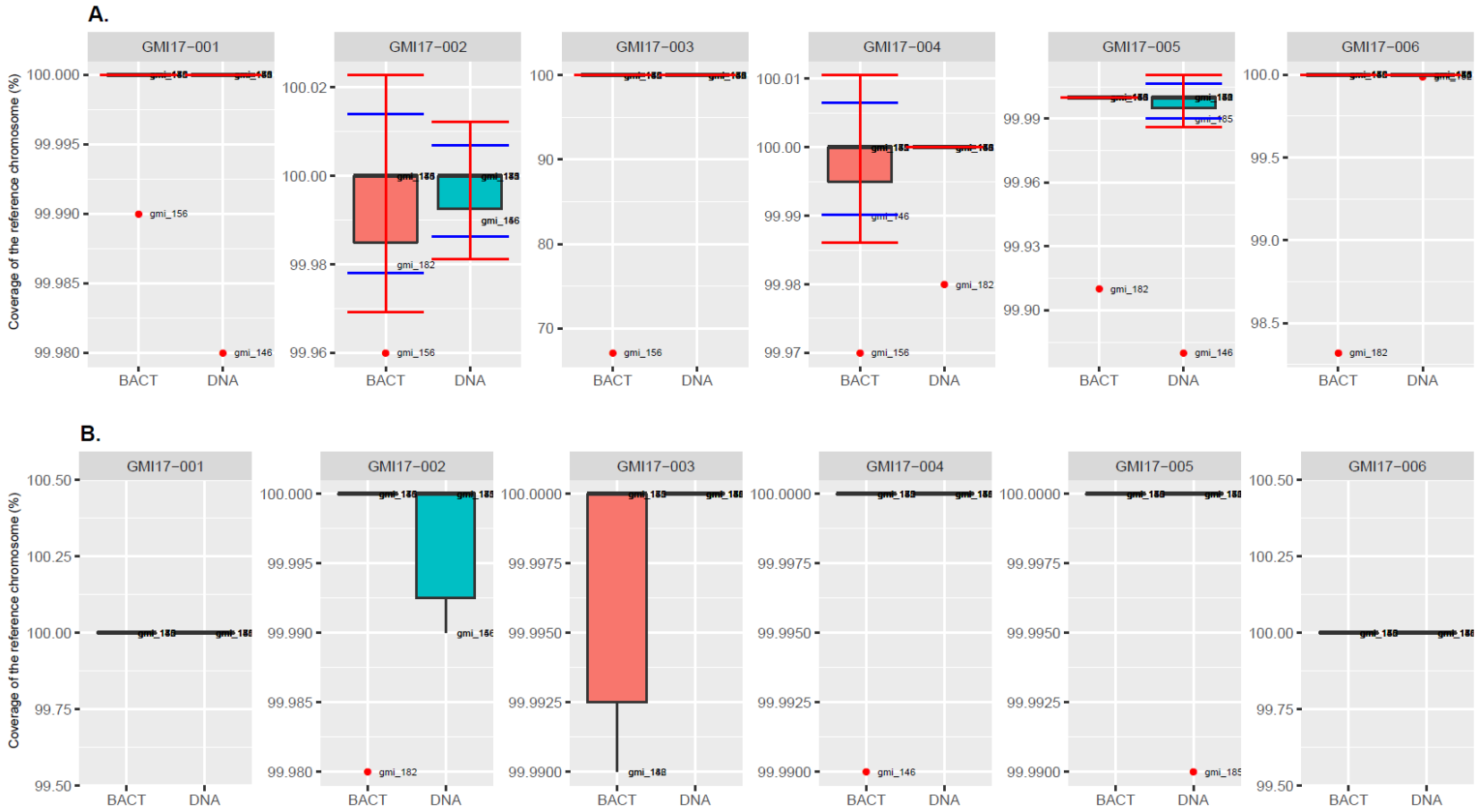


Figure M.21: Coverage of the reference chromosome

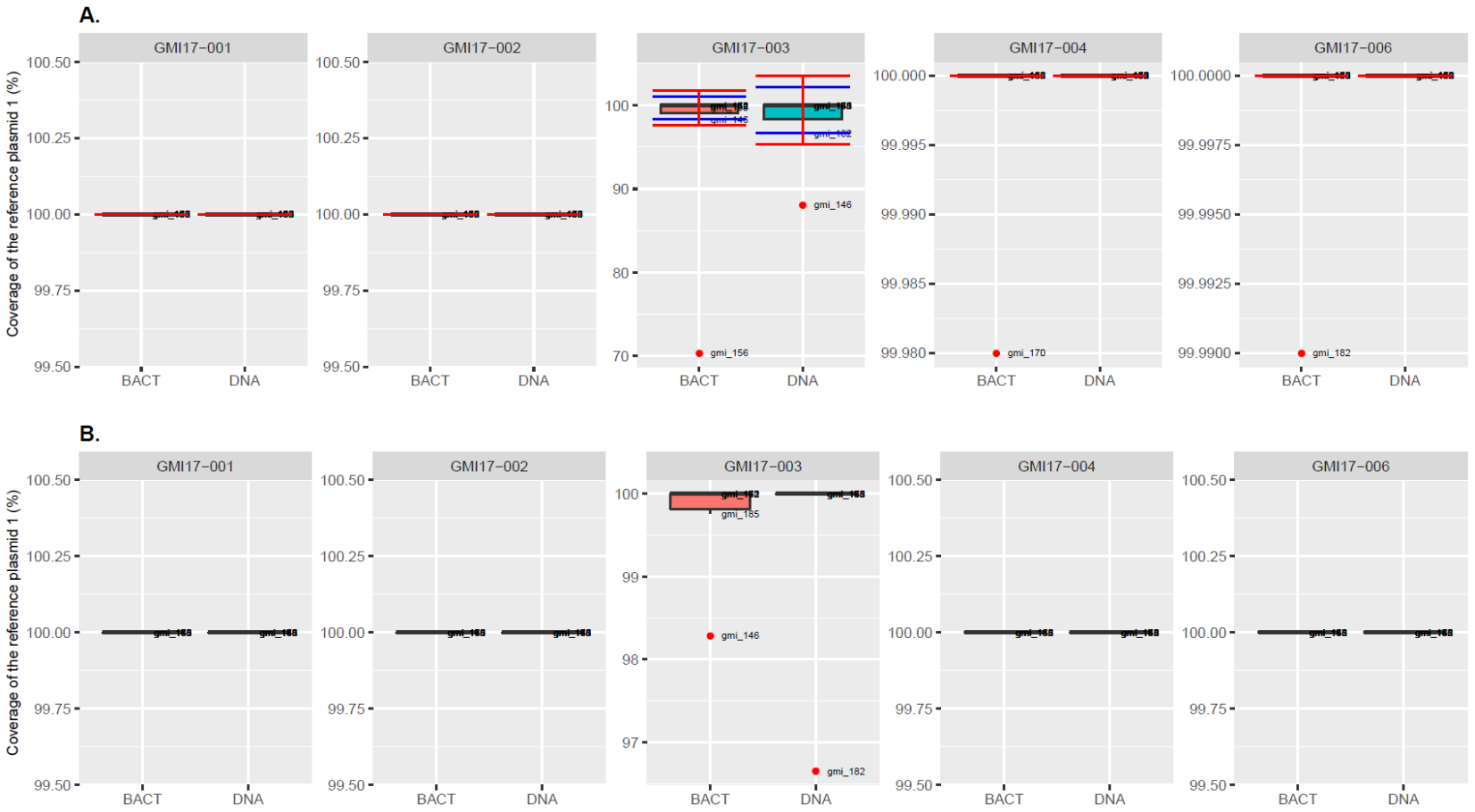


Figure M.22: Coverage of reference plasmid 1

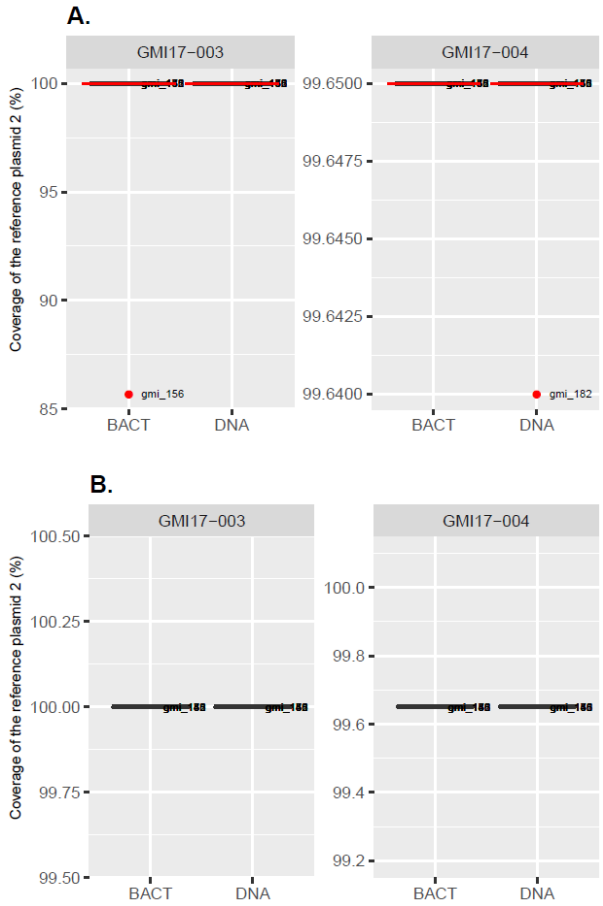


Figure M.23: Coverage of reference plasmid 2

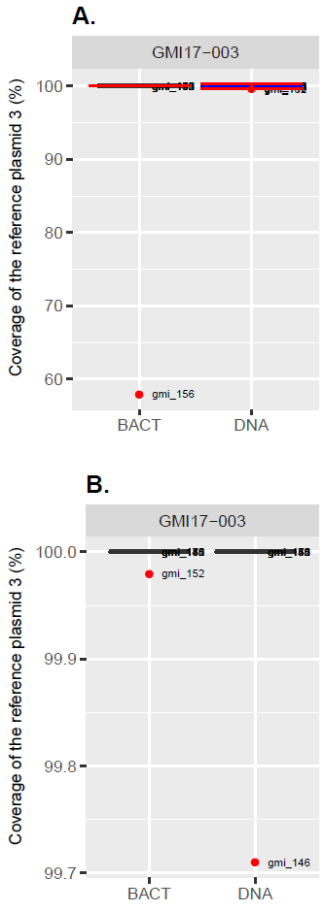


Figure M.24: Coverage of reference plasmid 3

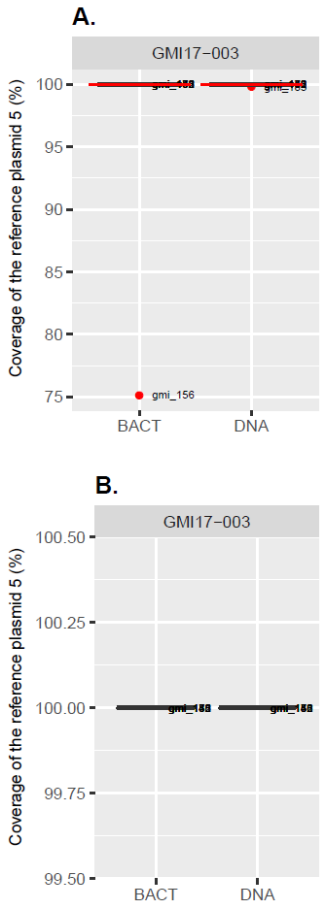


Figure M.25: Coverage of reference plasmid 4

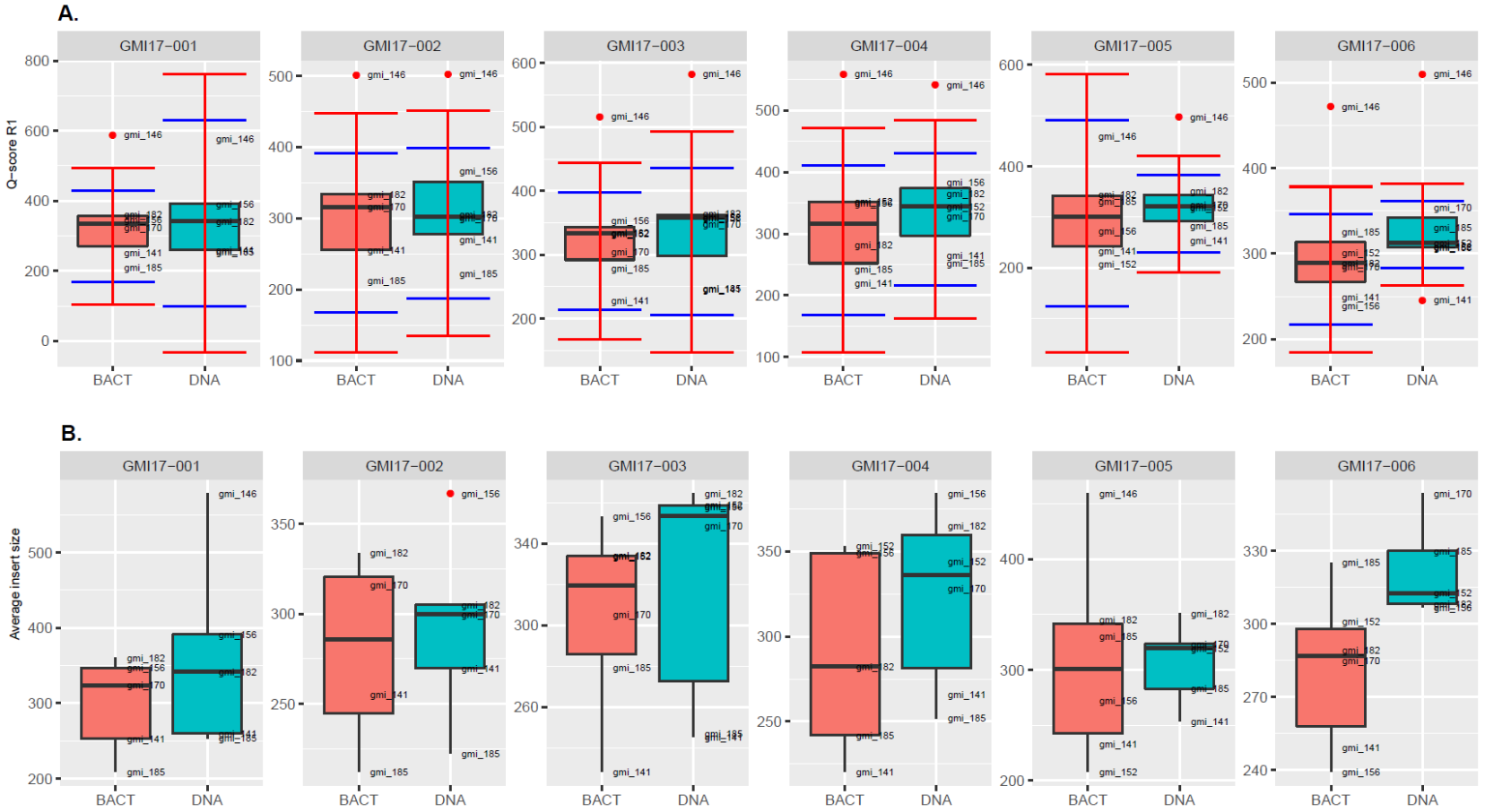


Figure M.26: Average insert size

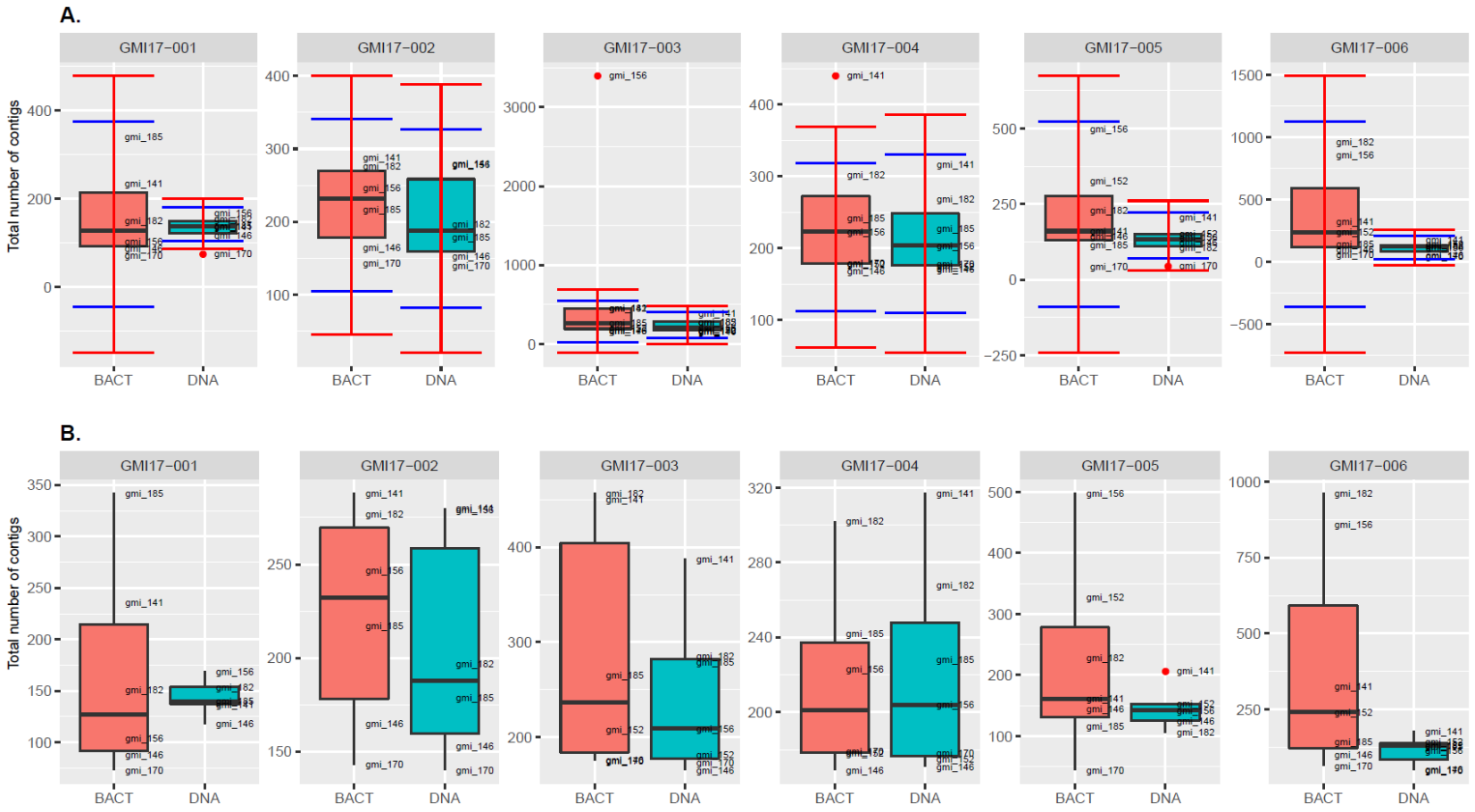


Figure M.27: Total number of contigs

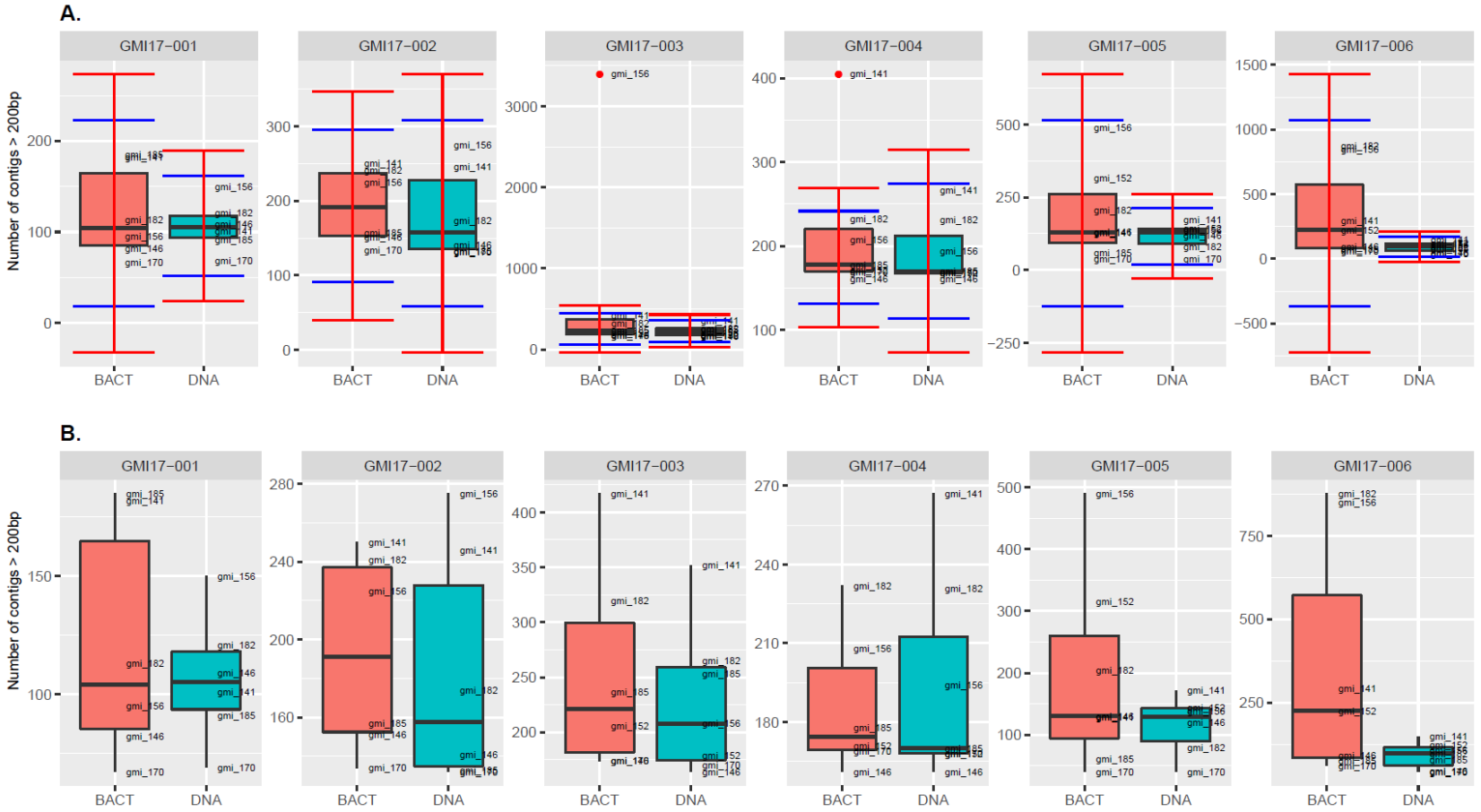


Figure M.28: Number of contigs > 200 bp

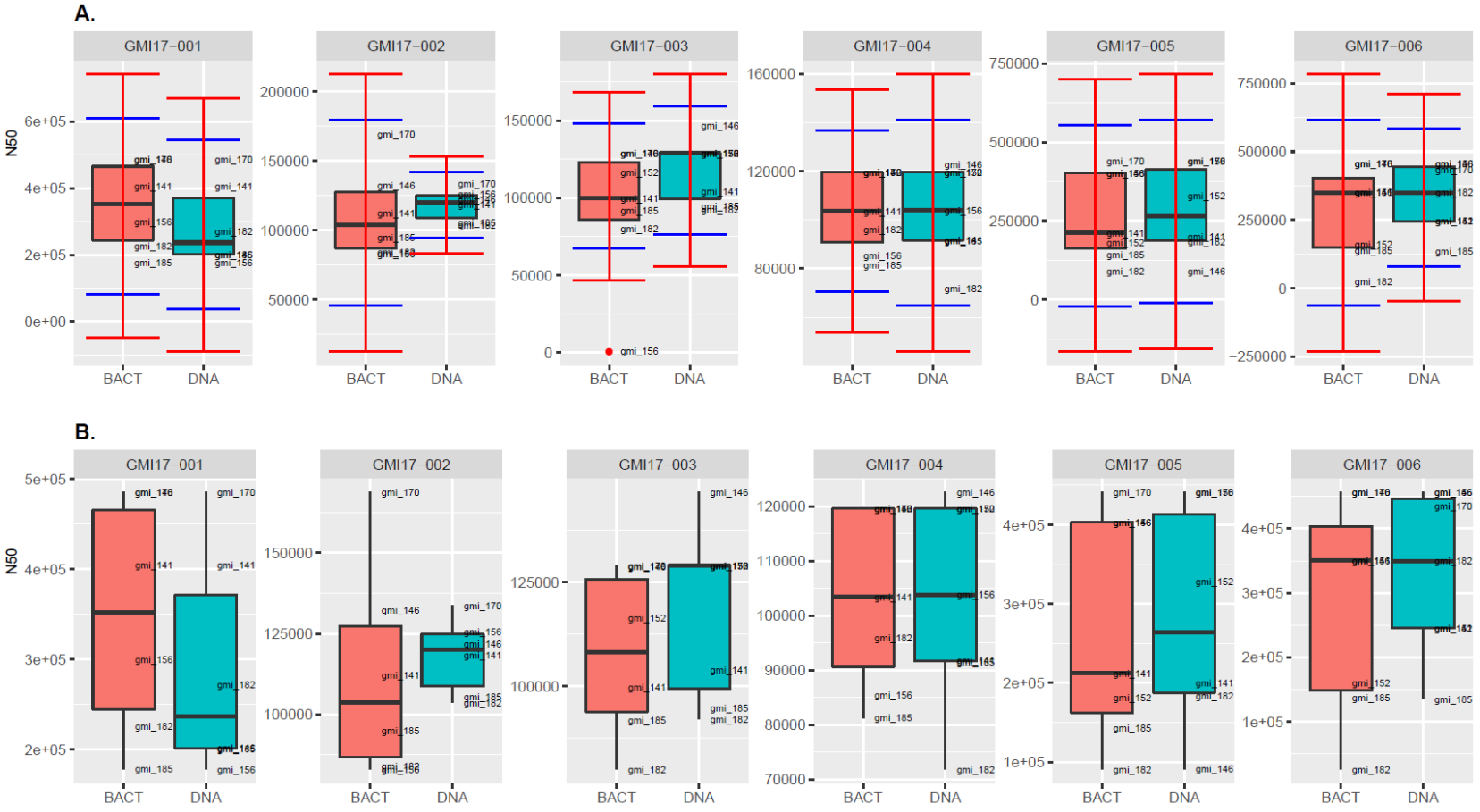


Figure M.29: N50